J. [illegible]losophy

Vol 82
1985

G. K. V.
Hardwar

110154

110254

RJ-1084

THE JOURNAL OF PHILOSOPHY

Vol 82

VOLUME LXXXII, NUMBER 1
JANUARY 1985

~~No 8~~ अगस्त
No 11 नवम्बर

page

5 *States of Mind* W. V. Quine
8 *Reduction, Qualia, and the Direct Introspection of Brain States* Paul M. Churchland

BOOK REVIEWS
28 Frank Jackson: *Perception: A Representative Theory* David M. Rosenthal
42 David Lewis: *Philosophical Papers* Christopher Peacocke
46 Anthony Kenny: *Faith and Reason*; J. L. Mackie: *The Miracle of Theism*; Terence Penelhum: *God and Skepticism*; Alvin Plantinga and Nicholas Wolterstorff, eds.: *Faith and Rationality* Richard Swinburne

54 MEMO TO AUTHORS
55 JOHNSONIAN PRIZE IN PHILOSOPHY
56 MEMO ON COPYRIGHT

193
Rec 20/4/85

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Editors: Bernard Berofsky, Arthur C. Danto, Hidé Ishiguro, Isaac Levi, Mary Mothersill, Sidney Morgenbesser, Charles D. Parsons, James J. Walsh. *Consulting Editor:* James Higginbotham. *Managing Editor:* Leigh S. Cauman. *Editorial & Business offices:* 720 Philosophy Hall, Columbia University, New York City 10027.

All communications to the editors, manuscripts, advertisements, subscriptions, and changes of address should be sent to the Managing Editor. Annual subscription price (12 numbers), to individuals, $15.00; to institutions, $20.00; to students and the retired and the unemployed, $12.00; current issues, $2.00; special issues, $3.00. All back volumes and separate issues are available back to 1904. Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, Pennsylvania.

THE JOURNAL OF PHILOSOPHY is owned and published by The Journal of Philosophy, Inc. *President,* Albert G. Redpath†; *Vice President,* Corliss Lamont; *Secretary,* Michael J. Mooney; *Treasurer,* Stephen S. Fenichell.

NOTE TO CONTRIBUTORS

As a monthly periodical, THE JOURNAL OF PHILOSOPHY encourages philosophical exchange and progress through prompt publication of articles of current interest. The editors now feel that they can ensure publication within four to eight months of acceptance. They continue to be particularly interested in papers that explore the borderline between philosophy and the special disciplines, but are unlikely to accept articles primarily expository or historical in character. Papers read at Eastern APA colloquia, which are abstracted in the JOURNAL, are not eligible for publication in full. If and only if articles are held five months or more and then rejected, letters of rejection include a brief comment concerning the editors' reasons for rejection.

Attention is drawn to the memo to authors and to the statement of copyright policy republished on pages 54 and 56 of this issue; copies are available from the JOURNAL office.

Contributors are supplied free of charge with ten copies of the JOURNAL issues in which their articles appear, reviewers and authors of abstracts with two copies. Offprints may be ordered from the printer on a form supplied by the printer and enclosed with galley proof; they will be sent out within six weeks of publication.

© Copyright 1985 by the Journal of Philosophy Inc.

ISSN 0022-362X

Second class postage paid at New York, N.Y., and at additional mailing offices.

POSTMASTER: Send address changes to 720 Philosophy Hall, Columbia U. NYC, NY 10027.

THE JOURNAL OF PHILOSOPHY

110154

VOLUME LXXXII, NUMBER 2
FEBRUARY 1985

193
20/4/85

page

57 *Probabilities and Causes* David Papineau
74 *Translucent Belief* Catherine Z. Elgin

COMMENTS AND CRITICISM

91 *Levi's "The Wrong Box"* Ellery Eells
104 *Epicycles* Isaac Levi

107 NOTES AND NEWS

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Editors: Bernard Berofsky, Arthur C. Danto, Hidé Ishiguro, Isaac Levi, Mary Mothersill, Sidney Morgenbesser, Charles D. Parsons, James J. Walsh. *Consulting Editor:* James Higginbotham. *Managing Editor:* Leigh S. Cauman. *Editorial & Business offices:* 720 Philosophy Hall, Columbia University, New York City 10027.

All communications to the editors, manuscripts, advertisements, subscriptions, and changes of address should be sent to the Managing Editor. Annual subscription price (12 numbers), to individuals, $15.00; to institutions, $20.00; to students and the retired and the unemployed, $12.00; current issues, $2.00; special issues, $3.00. All back volumes and separate issues are available back to 1904. Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, Pennsylvania.

THE JOURNAL OF PHILOSOPHY is owned and published by The Journal of Philosophy, Inc. *Acting President,* Corliss Lamont; *Secretary,* Michael J. Mooney; *Treasurer,* Stephen S. Fenichell.

The Philosophical Quarterly

Vol 34 No.137 October 1984

Edited in the University of St. Andrews by an Editorial Board
Chairman of the Board of Editors: LESLIE STEVENSON
Executive Editor: GORDON GRAHAM *Reviews Editor:* ANTHONY ELLIS

An Unsolved Puzzle about Knowledge	*Thomas Tymoczko*
'Ought implies can' and two kinds of morality	*John Kekes*
What does the private language argument prove?	*Oswald Hanfling*
Evolutionary Epistemology and Ontological Realism	*A.J. Clark*
DISCUSSIONS	
The Failure of an *A Priori* Argument for Realism	*Anthony L. Brueckner*
Ackerman on Propositional Identity	*Rod Bertolet*
Humes Dialogue IX Defended	*Donald E. Stahl*
Lucas, Godel and Astaire: a rejoinder	*J.R. Lucas*

Subscriptions for 1984

Annual Subscription	UK	Overseas	USA	Canada
Institutions	£22.50	£27.50	$50.00	$62.00
Individuals	£10.25	£12.50	$25.00	$30.00

Orders, remittances and enquiries should be addressed to: Journals Department,
Basil Blackwell Publisher Ltd., 108 Cowley Road, Oxford OX4 1JF, England.

The appearance of the code at the bottom of the first page of an article in this JOURNAL indicates the copyright owner's consent that copies of the article be made for personal or classroom use. This consent is given on condition, however, that the copier pay the stated per-copy fee through the Copyright Clearance Center, Inc., P.O. Box 765, Schenectady, New York 12301, for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying.

© Copyright 1985 by the Journal of Philosophy, Inc.

ISSN 0022-362X

Second class postage paid at New York, N.Y., and at additional mailing offices.
POSTMASTER: Send address changes to 720 Philosophy Hall, Columbia U, NYC, NY 10027.

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXII, NUMBER 3
MARCH 1985

पुस्तकालय गुरुकुल कांगड़ी
आगत संख्या 193
प्राप्ति दिनांक 2.5.85

page
113 *Moral Conflict and Moral Realism* Torbjörn Tännsjö
118 *Supererogation and Obligation* Frances Myrna Kamm

COMMENTS AND CRITICISM

139 *Moral Dilemmas, Deliberation, and Choice* Lyle V. Anderson
162 *A Solution to Forrester's Paradox of Gentle Murder* Walter Sinnott-Armstrong

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Editors: Bernard Berofsky, Arthur C. Danto, Hidé Ishiguro, Isaac Levi, Mary Mothersill, Sidney Morgenbesser, Charles D. Parsons, James J. Walsh. *Consulting Editor:* James Higginbotham. *Managing Editor:* Leigh S. Cauman. *Editorial & Business offices:* 720 Philosophy Hall, Columbia University, New York City 10027.

All communications to the editors, manuscripts, advertisements, subscriptions, and changes of address should be sent to the Managing Editor. Annual subscription price (12 numbers), to individuals, $15.00; to institutions, $20.00; to students and the retired and the unemployed, $12.00; current issues, $2.00; special issues, $3.00. All back volumes and separate issues are available back to 1904. Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, Pennsylvania.

THE JOURNAL OF PHILOSOPHY is owned and published by The Journal of Philosophy, Inc. *Acting President,* Corliss Lamont; *Secretary,* Michael J. Mooney; *Treasurer,* Stephen S. Fenichell.

REVUE INTERNATIONALE DE PHILOSOPHIE

Editor : Michel MEYER

143, av. A. Buyl, 1050 Brussels, Belgium

Each number is devoted to a particular movement, a particular philosopher, or a particular problem.

We publish 3 issues annually. Articles are written in English, French, German, or Italian.

Coming numbers :

- L'herméneutique
- Sartre, Common Sense
- Bacon
- Berkeley
- Platon

Our last issue was devoted to *Bachelard*

(articles by D. Bourel, Cl. Ramnoux, J. Starobinski, Fr. Dagognet, G.-G. Granger, J.-T. Desanti and J. G. Clark)

The appearance of the code at the bottom of the first page of an article in this JOURNAL indicates the copyright owner's consent that copies of the article be made for personal or classroom use. This consent is given on condition, however, that the copier pay the stated per-copy fee through the Copyright Clearance Center, Inc., P.O. Box 765, Schenectady, New York 12301, for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying.

© Copyright 1985 by the Journal of Philosophy Inc.

ISSN 0022-362X

Second class postage paid at New York, N.Y., and at additional mailing offices.

POSTMASTER: Send address changes to 720 Philosophy Hall, Columbia U. NYC, NY 10027.

THE JOURNAL OF PHILOSOPHY

THE JOURNAL OF PHILOSOPHY was founded in 1904 as THE JOURNAL OF PHILOSOPHY, PSYCHOLOGY AND SCIENTIFIC METHODS by Frederick J. E. Woodbridge and J. McKeen Cattell. In 1906 Wendell T. Bush became associated with the JOURNAL as co-editor. In 1923, the JOURNAL was incorporated in the State of New York under its present name. Since its founding it has been published from Columbia University and printed by the Lancaster Press in Lancaster, Pennsylvania.

THE JOURNAL OF
PHILOSOPHY

VOLUME LXXXII
JANUARY-DECEMBER 1985

Edited by
Bernard Berofsky, Arthur C. Danto
Hide Ishiguro, Isaac Levi, Mary Mothersill
Sidney Morgenbesser, Charles D. Parsons
James J. Walsh

Published by The Journal of Philosophy, Inc.

The appearance of the code at the bottom of the first page of an article in this JOURNAL indicates the copyright owner's consent that copies of the article be made for personal or classroom use. This consent is given on condition, however, that the copier pay the stated per-copy fee through the Copyright Clearance Center, Inc., P. O. Box 765, Schenectady, New York 12301, for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying.

© COPYRIGHT 1985 BY THE JOURNAL OF PHILOSOPHY, INC.

PRINTED BY LANCASTER PRESS, INC., LANCASTER, PA.

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXII, NO. 1, JANUARY 1985

STATES OF MIND*

UNLESS a case is to be made for disembodied spirits, we can argue that a dualism of mind and body is an idle redundancy. Corresponding to every mental state, however fleeting or however remotely intellectual, the dualist is bound to admit the existence of a bodily state that obtains when and only when the mental one obtains. The bodily state is trivially specifiable in the dualist's own terms, simply as the state of accompanying a mind that is in that mental state. Instead of ascribing the one state to the mind, then, we may equivalently ascribe the other to the body. The mind goes by the board, and will not be missed.

Not that we would continue to refer to the bodily state as the state of accompanying a mind in the mental state. That formulation was for the dualist's benefit, to show him that the bodily state was undeniable from his own point of view and specifiable in his own terms. For our part, we just appropriate the mentalistic terms themselves and construe them as referring to those bodily states. We even continue to speak of the states as mental. The only change is that we reckon mental states as states of the body rather than as states of another substance, the mind.

If this effortless physicalism smacks of trickery, we may do well to reflect on how we learn the mentalistic terms in the first place. All talk about one's mental life presupposes external reference. Introspect our mental states as we will, how do we know what to call them? How did we learn to call our anxieties anxieties, our dull aches dull aches, our joys joys and our awareness awareness? Why do we suppose that what we call joys and anxieties are what other people call by those names? How do we know what we are talking about? Clearly the answer is that such terms are applied in the

---

*This is an outgrowth of my piece in a colloquium of the American Philosophical Association, Detroit, 1980. Burton Dreben lately suggested some changes and urged publication. The key idea of the first half was briefly noted meanwhile in my *Theories and Things* (Cambridge, Mass.: Harvard, 1981), p. 18.

0022-362X/85/8201/0005$00.50

light of publicly observable symptoms: bodily symptoms strictly of bodily states, and the mind is as may be. Someone observes my joyful or anxious expression, or perhaps observes my gratifying or threatening situation itself, or hears me tell about it. She then applies the word 'joy' or 'anxiety'. After another such lesson or two I find myself applying these words to some of my subsequent states in cases where no outward signs are to be observed beyond my report itself. Without the outward signs to begin with, mentalistic terms could not be learned at all. To apply these terms to the state of the body is just to put them back where they belonged to begin with.

I am not applying the terms to behavior. A mental state is not always manifested in behavior. Physically construed, it is a state of nerves. We can say which state it is, however, and tell one from another, without knowing the neural mechanism. We specify it with help of the mentalistic term, which in turn was learned on the strength of behavioral signs. Thus the net result, if we shortcut the mental bit, is that the behavior provides incomplete and sporadic symptoms whereby to identify and distinguish various complex states of nerves whereof the neural detail may still be a matter of conjecture; and it is these states that the mental terms may be seen as denoting.

Mental states, construed as states of nerves, are like diseases. A disease may be diagnosed in the light of observable signs though the guilty germ be still unknown to science. Incidentally, diagnosis depends heavily on symptoms reported by the patient; and such is the way, overwhelmingly, with the detection of mental states.

What now can we make of the difference between identifying the mental states with the states of nerves, as I just did, and repudiating them rather in favor of states of nerves? I see no difference. In either case the states of nerves are retained, mental states in any other sense are repudiated, and the mental terms are thereupon appropriated to states of nerves. So I may as well persist in calling my proposed reduction of mind to body an identificaiton of mental states with bodily ones, neural ones; a construing of the mental as neural.

There is no presumption that the mentalistic idioms would in general be translatable into the anatomical and biochemical terminology of neurology, even if all details of the neurological mechanisms were understood. Thus take belief. Assessed on its objective manifestations, belief is a very mixed bag. Lip service is our most convenient clue to belief, but is neither a necessary nor a sufficient condition. Acceptance of wagers is a firmer sign, and the accepted odds even afford a measure of the strength of the belief; but this test is available only if there is a prospect of subsequently finding the answer, acceptably to both parties, and settling the bet. Other behavior, such as searching or fleeing or standing expectantly, can serve tentatively as manifestations of one or another belief, but these manifestations vary drastically and unsystematically with the content of the belief to be ascribed. Other grounds for ascribing beliefs may be sought unsystematically by probing the subject's past for probable causes of his present state of mind, or by seeing how he will defend his purported belief when challenged.

The empirical content of ascriptions of belief is thus heterogeneous in the extreme, and the physiological mechanisms involved are no less so. The heterogeneity is cloaked under a linguistic uniformity: the connective 'believes that' followed by a subordinate sentence. The other idioms of propositional attitude have the same disarmingly uniform structure as the belief idiom, and it cloaks much the same heterogeneity in respect of empirical evidence and neural mechanism. Hence the insistence, from Brentano onward, on the need for an independent science of intension.[1] Even those of us who do not acquiesce in a metaphysical dualism of mind and body must take the best of what Davidson has called anomalous monism.[2]

The stubborn idioms of propositional attitude are as deeply rooted as the overtly physicalistic ones. One of them is almost coeval, it would seem, with observation sentences. Thus take the observation sentence 'It's raining'. Tom is learning it from Martha by ostension. Martha's business is to encourage Tom in uttering the sentence, or in assenting to it, when she sees that he is noticing appropriate phenomena, and to discourage him otherwise. Thus Tom's mastery of the physicalistic sentence 'It's raining' hinges on Martha's mastery, virtual if not literal, of the mentalistic sentence 'Tom perceives that it's raining'.

Observation sentences, learned ostensively, are where our command of language begins, and our learning them from our elders depends heavily on the ability of our elders to guess that we are getting the appropriate perception. The handing down of language is thus implemented by a continuing command, tacit at least, of the idiom 'x perceives that p' where 'p' ranges over observation sentences. This degree of implicit mastery of one idiom of propositional attitude would seem therefore to be nearly as old as language. The structure of this idiom, moreover—its embedding of a subordinate sentence—would have been clearly dictated by its primitive use in assessing children's acquisition of observation sentences. Analogical extension of the idiom to other than observation sentences would follow inevitably, and the development of parallel idioms for other propositional attitudes would then come naturally too, notwithstanding their opacity from a logical point of view. Naturalness is one thing, transparency another; familiarity one, clarity another.

---

[1] See Roderick Chisholm, *Perceiving* (Ithaca, N.Y.: Cornell, 1957), chapter 11.
[2] Donald Davidson, *Action and Events* (New York: Oxford, 1980), pp. 214–225.

W. V. QUINE

Harvard University

REDUCTION, QUALIA, AND THE DIRECT INTROSPECTION OF BRAIN STATES*

DO the phenomenological or qualitative features of our sensations constitute a permanent barrier to the reductive aspirations of any materialistic neuroscience? I here argue that they do not. Specifically, I wish to address the recent anti-reductionist arguments posed by Thomas Nagel,[1] Frank Jackson,[2] and Howard Robinson.[3] And I wish to explore the possibility of human subjective consciousness within a conceptual environment constituted by a matured and successful neuroscience.

If we are to deal sensibly with the issues here at stake, we must approach them with a general theory of scientific reduction already in hand, a theory motivated by and adequate to the many instances and varieties of interconceptual reduction displayed *elsewhere* in our scientific history. With an independently based account of the nature and grounds of intertheoretic reduction, we can approach

---

*Research for this paper was supported by a grant from the Institute for Advanced Study, by a research/study leave from the University of Manitoba, and by SSHRC grant no. 451-83-3050. An earlier draft was presented to the Conference on Methods in Philosophy and the Sciences, convened in honor of Ernest Nagel, at the New School of Social Research in April, 1983. My thanks to that audience for their insightful criticisms and helpful suggestions. Thanks also to Thomas Nagel, Frank Jackson, Daniel Dennett, Philip Hanson, Charles Marks, and Brian Loar for critical discussions of that earlier draft.

[1] "What Is It Like to Be a Bat?", *Philosophical Review*, LXXXIII, 4 (October, 1974): 435–450; page references to Nagel are to this paper.

[2] "Epiphenomenal Qualia," *Philosophical Quarterly*, XXXII, 127 (April 1982): 127–136.

[3] *Matter and Sense* (New York: Cambridge, 1982), p. 4.

0022-362X/85/8201/0008$02.00

the specific case of subjective qualia, free from the myopia that results from trying to divine the proper conditions on reduction by simply staring long and hard at the problematic case at issue.

I. INTERTHEORETIC REDUCTION

We may begin by remarking that the classical account of intertheoretic reduction[4] now appears to be importantly mistaken, though the repairs necessary are quickly and cleanly made. Suppressing niceties, we may state the original account as follows. A new and more comprehensive theory *reduces* an older theory just in case the new theory, when conjoined with appropriate correspondence rules, logically entails the principles of the older theory. (The point of the correspondence rules, or "bridge laws," is to connect the disparate ontologies of the two theories: often these are expressed as identity statements, such as $Temperature = mv^2/3k$.) Schematically,

$$T_N \mathbin{\&} (\text{Correspondence Rules})$$

logically entails

$$T_O$$

Difficulties with this view begin with the observation that most reduced theories turn out to be, strictly speaking and in a variety of respects, *false*. (Real gases don't really obey $PV = \mu RT$, as in classical thermodynamics; the planets don't really move in ellipses, as in Keplerian astronomy; the acceleration of falling bodies isn't really uniform, as in Galilean dynamics; etc.) If reduction is *de*duction, modus tollens would thus require that the premises of the new reducing theories (statistical thermodynamics in the first case, Newtonian dynamics in the second and third) be somehow false as well, in contradiction to their assumed truth.

This complaint can be temporarily deflected by pointing out that the premises of a reduction must often include not just the new reducing theory but also some limiting assumptions or counterfactual boundary conditions (such as that the molecules of a gas enjoy only mechanical energy, or that the mass of the planets is negligible compared to the sun's, or that the distance any body falls is negligibly different from zero). Falsity in the reducing premises can thus be conceded, since it is safely confined to those limiting or counterfactual assumptions.

This defense will not deal with all cases of falsity, however, since in some cases the reduced theory is so radically false that some or all of its ontology must be rejected entirely, and the "correspondence rules" connecting that ontology to the newer ontology therefore display a problematic status. Newly conceived features cannot be identical with, nor even nomically connected with, old features, if the old features are illusory and uninstantiated. For example, relativistic mass is not identical with Newtonian mass, nor even coextensive with it, even at low velocities. Nevertheless, the reduction of Newtonian by Einsteinian mechanics is a paradigm of a successful reduction. For a second example, neither is caloric-fluid-pressure identical with, nor even coextensive with, mean molecular kinetic energy. But an overtly *fluid* thermodynamics (i.e., one committed to the existence of caloric) still finds a moderately impressive reduction within statistical thermodynamics. In sum, even theories with a *nonexistent* ontology can enjoy reduction, and this fact is problematic on the traditional account at issue.

---

[4] Ernest Nagel, *The Structures of Science* (New York: Harcourt, Brace & World, 1961), ch. 11.

What cases like these invite us to give up is the idea that what gets *de*duced in a reduction is the theory to be *re*duced. A more accurate, general, and illuminating schema for intertheoretic reduction is as follows:

T_N & Limiting Assumptions & Boundary Conditions

logically entails

I_N [a set of theorems of (restricted) T_N]
e.g., $(x)\,(Ax \supset Bx)$
$(x)\,((Bx \,\&\, Cx) \supset Dx)$

which is relevantly isomorphic with

T_O (the older theory)
e.g., $(x)\,(Jx \supset Kx)$
$(x)\,((Kx \,\&\, Lx) \supset Mx)$

That is to say, a reduction consists in the deduction, within T_N, not of T_O itself, but rather of a roughly equipotent *image* of T_O, an image still expressed in the vocabulary proper to T_N. The correspondence rules play no part whatever in the *de*duction. They show up only later, and not necessarily as material-mode statements, but as mere ordered pairs: $\langle Ax, Jx \rangle$, $\langle Bx, Kx \rangle$, $\langle Cx, Lx \rangle$, $\langle Dx, Mx \rangle$. Their function is to indicate which term substitutions in the image I_N will yield the principles of T_O. The older theory, accordingly, is never deduced; it is just the target of a relevantly adequate *mimicry*. Construed in this way, a correspondence rule is entirely consistent with the assumption that the older predicate it encompasses has no extension whatever. This allows that a true theory might reduce even a substantially false theory.

The point of a reduction, according to this view, is to show that the new or more comprehensive theory contains explanatory and predictive resources that parallel, to a relevant degree of exactness, the explanatory and predictive resources of the reduced theory. The *intra*-theoretic deduction (of I_N within T_N), and the *inter*-theoretic mapping (of T_O into I_N), constitute a fell-swoop demonstration that the older theory can be displaced wholesale by the new, without significant explanatory or predictive loss.[5]

Material-mode statements of identity can occasionally be made, of course. We do wish to assert that visible light = EM waves between .35 μm and .75 μm, that sound = atmospheric compression waves, that temperature = mean molecular KE, and that electric current = net motion of charged particles. But a correspondence rule does not itself make such a claim: at best, it records the fact that the new predicate applies in all those cases where its T_O-doppelganger predicate was normally *thought* to apply. On this view, full-fledged *identity* statements are licensed by the comparative *smoothness* of the relevant reduction (i.e., the limiting assumptions or boundary conditions on T_N are not wildly counterfactual, all or most of T_O's principles find close analogues in I_N, etc.). This smoothness permits the comfortable assimilation of the old ontology within the new and thus allows the old theory to retain all or most of its ontological integrity. *It is smooth intertheoretic reductions that motivate and sustain statements of cross-theoretic identity, not the other way around.*

The preceding framework allows us to frame a useful conception of reduction for specific *properties*, as opposed to entire theories, and it allows us to frame a useful conception of the contrary notion of "emergent" properties. A property F, postulated by an older theory or conceptual framework T_O, is reduced to a property G in some new theory T_N just in case

(1) T_N reduces T_O;
(2) F and G are correspondence-rule paired in the reduction; and
(3) The reduction is sufficiently smooth to sustain the ontology of T_O, and thus to sustain the identity claim, "F-ness = G-ness."

Intuitively, and in the material mode, this means that F-ness reduces to G-ness just in case the "causal powers" of F-ness (as outlined in the laws of T_O) are a subset of the "causal powers" of G-ness (as outlined in the laws of T_N).

---

[5] This sketch of intertheoretic reduction is drawn from my *Scientific Realism and the Plasticity of Mind* (New York: Cambridge, 1979), section 11. For a more detailed account see Clifford A. Hooker, "Towards a General Theory of Reduction", *Dialogue*, xx, 1, 2, 3 (March, June, September 1981): 38-59, 201-236, 496-529.

Finally, a property *F* will be said to be an *emergent* property (relative to T_N) just in case

(1) *F* is definitely real and instantiated;
(2) *F* is co-occurrent with some feature or complex circumstance recognized in T_N; but
(3) *F* cannot be *reduced* to any property postulated by or definable within T_N.

Intuitively, this will happen when T_N does not have the resources adequate to define a property with all the "causal powers" possessed by *F*-ness. Claims about the emergence of certain properties are therefore claims about the relative poverty in the resources of certain aspirant theories.[6] Having outlined these notions, we shall turn to address substantive questions of emergence and irreducibility in a few moments.

Before we do so, several points about reduction need to be emphasized. The first is that, in arguing for the emergence of a given property *F* relative to some theory T_N, it is not sufficient to point out that the existence or appearance of *F*-ness cannot be deduced from T_N. It is occasionally claimed, for example, that the objective features of warmth or blueness must be irreducibly emergent properties, since, however much one bends and squeezes the molecular theory concerning H_2O, one cannot deduce from it that water will be *blue*, but only that water will scatter electromagnetic radiation at such-and-such wavelengths. And however much one wrings from the mechanics of molecular motion, one cannot deduce from it that a roaring hearth will be warm, but only that its molecules will have such-and-such a mean kinetic energy and will collectively emit EM radiation at longish wavelengths.

These premises about nondeducibility are entirely true, but the conclusion against reducibility does not follow. It is a serious mistake even to make *in*direct deducibility (i.e., deducibility with the help of correspondence rules) a requirement on successful reduction, as we saw at the beginning of this section. And there are additional reasons why it would be even more foolish to insist on the much stronger condition of direct deducibility. For example, formal considerations alone guarantee that, for any predicate '*F*' not already in the proprietary lexicon of the aspirant reducing theory T_N, no statements whatever involving '*F*' (beyond tautologies and other trivial exceptions) will be deducible from T_N. The deducibility requirement would thus trivialize the notion of reduction by making it impossible for *any* conceptual framework to reduce any other, distinct conceptual framework. Even temperature—that paradigm of a successfully reduced property—would be rendered irreducible, since the term 'temperature' does not appear in the lexicon of statistical mechanics.

---

[6] A word of caution is perhaps in order here, since the expression 'emergent property' is often used in two diametrically opposed senses. In scientific contexts, one frequently hears it used to apply to what might be called a "network property," a property that appears exactly when the elements of some substrate are suitably organized, a property that *consists in* the elements of that substrate standing in certain relations to one another, a set of relations that collectively sustain the set of causal powers ascribed to the "emergent" property. In this innocent sense of 'emergent', there are a great many emergent properties, and quite probably the qualia of our sensations should be numbered among them. But in philosophical contexts one more often encounters a different sense of 'emergent', one that implies that an "emergent" property does *not* consist in any collective or organizational feature of its substrate. The first sense positively implies reducibility; the second implies *irre*ducibility. It is emergence in the second sense that is at issue in this paper.

There is a further reason why the demand for direct deducibility is too strong. It is a historical accident that we humans currently use precisely the conceptual framework we do use. We might have used any one of an infinite number of other conceptual frameworks to describe the observable world, each one of which could have been roughly adequate to common experience and many of which would be roughly isomorphic (each in its different way) with some part of the correct account that a utopian theory will eventually provide. Accordingly, we can legitimately ask of a putatively correct theory of a given objective domain that it account for the phenomena in (= function successfully in) that domain. But we cannot insist that it also be able to predict how this, that, or the other conceptually idiosyncratic human culture is going to *conceive* of that domain. That would be to insist that the new theory do *predictive cultural anthropology* for us, as well as mechanics, or electromagnetic theory, or what have you. The demand that molecular theory directly entail *our* thermal or color concepts is evidently this same unreasonable demand.

All we can properly ask of a reducing theory is that it have the resources to conjure up a set of properties whose nomological powers/roles/features are systematic *analogues* of the powers/roles/features of the set of properties postulated by the old theory. Since both theories presume to describe the same empirical domain, these systematic nomological parallels constitute the best grounds there can be for concluding that both theories have managed to latch onto the *same* set of objective properties. The hypothesized identity of the properties at issue explains why I_N and T_O are taxonomically and nomically parallel: they are both at least partially correct accounts of the very same objective properties. I_N merely frames that account within a much more penetrating conceptual system—that of T_N.

Moreover, it is to be expected that existing conceptual frameworks will eventually be reduced or displaced by new and better ones, and those in turn by frameworks better still; for who will be so brash as to assert that the feeble conceptual achievements of our adolescent species comprise an exhaustive account of anything at all? If we put aside this conceit, then the only alternatives to intertheoretic reduction are epistemic stagnation or the outright elimination of old frameworks as wholly false and illusory.

II. THEORETICAL CHANGE AND PERCEPTUAL CHANGE

Esoteric properties and arcane theoretical frameworks are not the only things that occasionally enjoy intertheoretic reduction. Observable properties and common-sense conceptual frameworks can also enjoy smooth reduction. Thus, being a middle-A sound is identical with being an oscillation in air pressure at 440 hz; being red is identical with having a certain triplet of electromagnetic reflectance efficiencies; being warm is identical with having a certain mean level of microscopically embodied energies, and so forth.

Moreover, the relevant reducing theory is capable of replacing the old framework not just in contexts of calculation and inference. *It should be appreciated that the reducing theory can displace the old framework in all its observational contexts as well.* Given the reality of the property identities just listed, it is quite open to us to begin framing our spontaneous perceptual reports in the language of the more sophisticated reducing theory. It is even desirable that we begin doing this, since the new vocabulary observes distinctions that are in fact within the discriminatory reach of our native perceptual systems, though those objective distinctions go unmarked and unnoticed from within the old framework. We can thus make more penetrating use of our native perceptual equipment. Such displacement is also desirable for a second reason: the greater inferential or computational power of the new conceptual framework. We can thus make better inferential *use* of our new perceptual judgments than we made of our old ones.

It is difficult to convey in words the vastness of such perceptual transformations and the naturalness of the new conceptual regime, once established. A nonscientific example may help to get the initial point across.

Consider the enormous increase in discriminatory skill that spans the gap between an untrained child's auditory apprehension of a symphony and the same person's apprehension of the same symphony forty years later, heard in his capacity as conductor of the orchestra performing it. What was before a seamless voice is now a mosaic of distinguishable elements. What was before a dimly apprehended tune is now a rationally structured sequence of distinguishable and identifiable chords supporting an appropriately related melody line. The matured musician hears an entire world of structured detail, concerning which the child is both dumb and deaf.

Other modalities provide comparable examples. Consider the practiced and chemically sophisticated wine taster, for whom the "red wine" classification used by most of us divides into a network of fifteen or twenty distinguishable elements: ethanol, glycol, fructose, sucrose, tannin, acid, carbon dioxide, and so forth, whose relative concentrations he can estimate with accuracy.

Or consider the astronomer, for whom the speckled black dome of her youth has become a visible abyss, scattering nearby planets, yellow dwarf stars, blue and red giants, distant globular clusters, and even a remote galaxy or two, all discriminable as such and locatable in three-dimensional space with her unaided (repeat: *unaided*) eye.

In each of these cases, what is finally mastered is a conceptual framework—whether musical, chemical, or astronomical—a framework that embodies far more wisdom about the relevant sensory domain than is immediately apparent to untutored discrimination. Such frameworks are characteristically a cultural heritage, pieced together over many generations, and their mastery supplies a richness and penetration to our sensory lives that would be impossible in their absence.[7]

Our *introspective* lives are already the extensive beneficiaries of this phenomenon. The introspective discriminations we make are for the most part learned; they are acquired with practice and experience, often quite slowly. And the specific discriminations we learn to make are those it is useful for us to make. Generally, those are the discriminations that others are already making, the discriminations embodied in the psychological vocabulary of the language we learn. The conceptual framework for psychological states that is embedded in ordinary language is a modestly sophisticated theoretical achievement in its own right, and it shapes our matured introspection profoundly. If it embodied substantially *less* wisdom in its categories and connecting generalizations, our introspective apprehension of our internal states and activities would be much diminished, though our native discriminatory mechanisms remain the same. Correlatively, if folk psychology embodied substantially *more* wisdom about our inner nature than it actually does, our introspective discrimination and recognition could be very much *greater* than it is, though our native discriminatory mechanisms remain unchanged.

---

[7] The role of theory in perception, and the systematic enhancement of perception through theoretical progress, are examined at length in my *Scientific Realism and the Plasticity of Mind, op. cit.* secs. 1–6.

This brings me to the central positive suggestion of this paper. Consider now the possibility of learning to describe, conceive, and introspectively apprehend the teeming intricacies of our inner lives within the conceptual framework of a matured neuroscience, a neuroscience that successfully reduces, either smoothly or roughly, our common-sense folk psychology. Suppose we trained our native mechanisms to make a new and more detailed set of discriminations, a set that corresponded not to the primitive psychological taxonomy of ordinary language, but to some more penetrating taxonomy of states drawn from a completed neuroscience. And suppose we trained ourselves to respond to that reconfigured discriminative activity with judgments that were framed, as a matter of course, in the appropriate concepts from neuroscience.[8]

If the examples of the symphony conductor (who can hear the A*m*7 chords), the oenologist (who can see and taste the glycol), and the astronomer (who can see the temperature of a blue giant star) provide a fair parallel, then the enhancement in our introspective vision could approximate a revelation. Dopamine levels in the limbic system, the spiking frequencies in specific neural pathways, resonances in the *n*th layer of the occipital cortex, inhibitory feedback to the lateral geniculate nucleus, and countless other neurophysical niceties could be moved into the objective focus of our introspective discrimination, just as G*m*7 chords and A*dim* chords are moved into the objective focus of a trained musician's auditory discrimination. We will of course have to *learn* the conceptual framework of a matured neuroscience in order to pull this off. And we will have to *practice* its noninferential application. But that seems a small price to pay for the quantum leap in self-apprehension.

---

[8] I believe it was Paul K. Feyerabend and Richard Rorty who first identified and explored this suggestion. See Feyerabend, "Materialism and the Mind-Body Problem," *Review of Metaphysics*, XVII.1, 65 (September 1963); 49-66; and Rorty, "Mind-Body Identity, Privacy, and Categories", *ibid.* XIX.1, 73 (September 1965): 24-54. This occurred in a theoretical environment prepared largely by Wilfrid Sellars in "Empiricism and the Philosophy of Mind," in Herbert Feigl and Michael Scriven, eds., *Minnesota Studies in the Philosophy of Mind*, vol. I (Minneapolis: of Minnesota Press, 1956): secs. 45-63. The idea has been explored more recently in my "Eliminative Materialism and the Propositional Attitudes," this JOURNAL, LXXVIII, 2 (February 1981): 67-90.

All of this suggests that there is no problem at all in conceiving the eventual reduction of mental states and properties to neurophysiological states and properties. A matured and successful neuroscience need only include, or prove able to define, a taxonomy of kinds with a set of embedding laws that faithfully mimics the taxonomy and causal generalizations of *folk* psychology. Whether future neuroscientific theories will prove able to do this is a wholly empirical question, not to be settled a priori. The evidence for a positive answer is substantial and familiar, centering on the growing explanatory success of the several neurosciences.

But there is negative evidence as well: I have even urged some of it myself ("Eliminative Materialism and the Propositional Attitudes," *op. cit.*). My negative arguments there center on the explanatory and predictive poverty of folk psychology, and they question whether it has the categorial integrity to *merit* the reductive preservation of its familiar ontology. That line suggests substantial revision or outright elimination as the eventual fate of our mentalistic ontology. The qualia-based arguments of Nagel, Jackson, and Robinson, however, take a quite different line. They find no fault with folk psychology. Their concern is with the explanatory and descriptive poverty of any possible *neuroscience,* and their line suggests that emergence is the correct story for our mentalistic ontology. Let us now examine their arguments.

III. THOMAS NAGEL'S ARGUMENTS

For Thomas Nagel, it is the phenomological features of our experiences, the properties or *qualia* displayed by our sensations, that constitute a problem for the reductive aspirations of any materialistic neuroscience. In his classic position paper (*op. cit.*) I find three distinct arguments in support of the view that such properties will never find any plausible or adequate reduction within the framework of a matured neuroscience. All three arguments are beguiling, but all three, I shall argue, are unsound.

First Argument: What makes the proposed reduction of mental phenomena different from reductions elsewhere in science, says Nagel, is that

> It is impossible to exclude the phenomenological features of experience from a reduction, in the same way that one excludes the phenomenal features of an ordinary substance from a physical or chemical reduction of it—namely, by explaining them as effects on the minds of human observers (437).

The reason it is impossible to exclude them, continues Nagel, is that the phenomenological features are essential to experience and to the subjective point of view. But this is not what interests me about this argument. What interests me is the claim that reductions of various substances elsewhere in science *exclude the phenomenal features of the substance.*

This is simply false, and the point is extremely important. The phenomenal features at issue are those such as the objective redness of an apple, the warmth of a coffee cup, and the pitch of a sound. These properties are not excluded from our reductions. Redness, an objective phenomenal property of apples, is identical with a certain wavelength triplet of electromagnetic reflectance efficiencies. Warmth, an objective phenomenal property of objects, is identical with the mean level of the objects' microscopically embodied energies. Pitch, an objective phenomenal property of a sound, is identical with its oscillatory frequency. These electromagnetic and micromechanical properties, out there in the objective world, are genuine phenomenal properties. Despite widespread ignorance of their dynamical and microphysical details, it is these objective physical properties to which everyone's perceptual mechanisms are keyed.

The reductions whose existence Nagel denies are in fact so complete that one can already displace entirely large chunks of our common-sense vocabulary for observable properties and learn to frame one's perceptual judgments directly in terms of the reducing theory. The mean KE of the molecules in this room, for example, is currently about . . . 6.2×10^{-21} joules. The oscillatory frequency of this sound (I here whistle C one octave above middle C) is about 524 hz. And the three critical electromagnetic reflectance efficiencies (at .45, .53, and .63 μm) of this (white) piece of paper are all above 80 per cent. These microphysical and electromagnetic properties can be felt, heard, and seen, respectively. Our native sensory mechanisms can easily discriminate such properties, one from another, and their presence from their absence. They have been doing so for millennia. The "resolution" of these mechanisms is inadequate, of course, to reveal the microphysical details and the extended causal roles of the properties thus discriminated. But they are abundantly adequate to permit the reliable discrimination of the properties at issue.[9]

On this view, the standard perceptual properties are not "secondary" properties at all, in the standard sense which implies that they have no real existence save *inside* the minds of human observers. On the contrary, they are as objective as you please, with a wide variety of objective causal properties. Moreover, it would be a mistake even to try to "kick the phenomenal properties inwards," since that would only postpone the problem of reckoning their place in nature. We would only confront them again later, as we address the place in nature of mental phenomena. And, as Nagel correctly points out, the relocation dodge is no longer open to us, once the problematic properties are already located within the mind.

---

[9] See again my *Scientific Realism and the Plasticity of Mind, op. cit.*, secs. 2–6. See also Paul and Patricia Churchland, "Functionalism, Qualia, and Intentionality", *Philosophical Topics*, XII, 1 (October 1981): 121–145. Reprinted in J. I. Biro and R. W. Shahan, eds., in *Mind, Brain, and Function* (Norman: U of Oklahoma Press, 1982): 121–145.

Nagel concludes from this that subjective qualia are unique in being immune from the sort of reductions found elsewhere in science. I draw a very different conclusion. The *objective* qualia (redness, warmth, etc.) should never have been "kicked inwards to the minds of observers" in the first place. They should be confronted squarely, and they should be reduced where they stand: *out*side the human observer. As we saw, this can and has in fact been done. If objective phenomenal properties are so treated, then *subjective* qualia can be confronted with parallel forthrightness, and can be reduced where *they* stand: *in*side the human observer. So far then, the external and the internal case are not different: they are parallel after all.

Second Argument: A second argument urges the point that the intrinsic character of experiences, the qualia of sensations, are essentially accessible from only a single point of view, the subjective point of view of the experiencing subject. The properties of physical brain states, by contrast, are accessible from a variety of entirely objective points of view. We cannot hope adequately to account for the former, therefore, in terms of properties appropriate to the latter domain (cf. Nagel, 442–444).

This somewhat diffuse argument appears to be an instance of the following argument:

(1) The qualia of my sensations are directly known by me, by introspection, as elements of my conscious self.
(2) The properties of my brain states are *not* directly known by me, by introspection, as elements of my conscious self.
∴(3) The qualia of my sensations ≠ the properties of my brain states.

And perhaps there is a second argument here as well, a complement to the first:

(1) The properties of my brain states are known-by-the-various-external-senses, as having such-and-such physical properties.
(2) The qualia of my sensations are *not* known-by-the-various-external-senses, as having such-and-such physical properties.
∴(3) The qualia of my sensations ≠ the properties of my brain states.

The argument form here is apparently

(1) Fa
(2) $\sim Fb$
$\therefore$(3) $a \neq b$

Given Leibniz's law and the extensional nature of the property F, this is a valid argument form. But, in the examples at issue, F is obviously not an extensional property. The fallacy committed in both cases is amply illustrated in the following parallel arguments.

(1) Hitler is widely recognized as a mass murderer.
(2) Adolf Schicklgruber is *not* widely recognized as a mass murderer.
∴(3) Hitler ≠ Adolf Schicklgruber.

or

(1) Aspirin is known by John to be a pain reliever.
(2) Acetylsalicylic acid is *not* known by John to be a pain reliever.
∴(3) Aspirin ≠ acetylsalicylic acid.

or, to cite an example very close to the case at issue,

(1) Temperature is known by me, by tactile sensing, as a feature of material objects.
(2) Mean molecular kinetic energy is *not* known by me, by tactile sensing, as a feature of material objects.
∴(3) Temperature ≠ mean molecular kinetic energy.

The problem with all these arguments is that the "property" ascribed in premise 1 and withheld in premise 2 consists only in the subject item's being *recogized, perceived,* or *known* as something, *under some specific description or other.* Such apprehension is not a genuine feature of the item itself, fit for divining identities, since one and the same subject may be successfully recognized under one description (e.g., "qualia of my mental state"), and yet fail to be recognized under another, equally accurate, coreferential description (e.g., "property of my brain state"). In logician's terms, the propositional function:

x is known (perceived, recongized) by me, as an F

is one of a large number of *intensional contexts* whose distinguishing feature is that they do not always retain the same truth value through substitution of a coreferential or coextensive term for whatever holds the place of 'x'. Accordingly, that such a context (i.e., the one at issue) should show a difference in truth value for two terms 'a' and 'b' (i.e., 'qualia of my sensations' and 'property of my brain-states') is therefore hardly grounds for concluding that 'a' and 'b' cannot be coreferential or coextensive terms![10]

This objection is decisive, I think, but it does not apply to a different version of the argument, which we must also consider. It may be urged that one's brain states are more than merely not (yet) known by introspection: they are not know*able* by introspection under any circumstances. In correspondence, Thomas Nagel has advised me that what he wishes to defend is the following *modalized* version of the argument:

(1) My mental states are knowable by me by introspection.
(2) My brain states are *not* knowable by me by introspection.
∴(3) My mental states ≠ my brain states.

110154

Here Nagel will insist that being knowable-by-me-by-introspection is a genuine relational property of a thing and that this version of the argument is free of the intensional fallacy discussed above.

And so it is. But now the reductionist is in a position to insist that the argument contains a false premise: premise 2. At the very least, he can insist that (2) begs the question. For if mental states are indeed identical with brain states, then it is really brain states that we have been introspecting all along, though without appreciating their fine-grained nature. And if we can learn to think of and recognize those states under their familiar mentalistic descriptions—*as all of us have*—then we can certainly learn to think of and recognize them under their more pentrating neurophysiological descriptions. Brain states, that is, are indeed know*able* by introspection, and Nagel's argument commits the same error instanced below.

(1) Temperature is knowable by tactile sensing.
(2) Mean molecular kinetic energy is *not* knowable by tactile sensing.
∴(3) Temperature ≠ mean molecular kinetic energy.

Here the conclusion is known to be false. Temperature is indeed mean molecular kinetic energy. Since the argument is valid, it must therefore have a false premise. Premise 2 is clearly the stinker. Just as one can learn to feel that the summer air is about 70°F, or 21°C, so one can learn to feel that the mean KE of its molecules is about 6.2×10^{-21} joules; for, whether we realize it or not, that is the property our native discriminatory mechanisms are keyed to. And if one can come to know, by feeling, the mean KE of atmospheric molecules, why is it unthinkable that one might come to know, by introspection, the states of one's brain? (What would that feel like? It would feel exactly the same as introspecting the states of one's

---

[10] I believe it was Richard Brandt and Jaegwon Kim who first identified this fallacy specifically in connection with the identity theory, in "The Logic of the Identity Theory," this JOURNAL, LXIV, 17 (September 1967): 515-537.

mind, since they are one and the same states. One would simply employ a different and more penetrating conceptual framework in their description.)

One must be careful, in evaluating the plausibility of Nagel's second premise, to distinguish it from the second premise of the very first version of the argument, the version that commits the intensional fallacy. My guess is that Nagel has profited somewhat from the ambiguity here. For, in the first version, both premises are true. And in the second version, the argument is valid. Neither version, however, meets both conditions.

The matter of introspecting one's brain states will arise once more in the final section of this paper. For now, let us move on.

Third Argument: The last argument here is the one most widely associated with Nagel's paper. The leading example is the (mooted) character of the experiences enjoyed by an alien creature such as a bat. The claim is that, no matter how much one knew about the bat's neurophysiology and its interaction with the physical world, one could still not know, nor perhaps even imagine, what it is like to be a bat. Even total knowledge of the physical details still leaves something out. The lesson drawn is that the reductive aspirations of neurophysiology are doomed to dash themselves, unrealized, against the impenetrable keep of subjective qualia (cf. Nagel, 438 ff.).

This argument is almost identical with an argument put forward in a recent paper by Frank Jackson.[11] Since Jackson's version deals directly with humans, I shall confront the problem as he formulates it.

IV. JACKSON'S KNOWLEDGE ARGUMENT

Imagine a brilliant neuroscientist named Mary, who has lived her entire life in a room that is rigorously controlled to display only various shades of black, white, and grey. She learns about the outside world by means of a black/white television monitor, and, being brilliant, she manages to transcend these obstacles. She becomes the world's greatest neuroscientist, all from within this room. In particular, she comes to know everything there is to know about the physical structure and activity of the brain and its visual system, of its actual and possible states.

But there would still be something she did *not* know, and could not even imagine, about the actual experiences of all the other people who live outside her black/white room, and about her possible experiences were she finally to leave her room: the nature of the experience of seeing a ripe tomato, what it is like to see red or have a sensation-of-red. Therefore, complete knowledge of the physical facts of visual perception and its related brain activity *still leaves something out.* Therefore, materialism cannot give an adequate reductionist account of all mental phenomena.

---

[11] "Epiphenomenal Qualia," *op. cit.* Howard Robinson runs a very similar argument in *Matter and Sense, op. cit.*, p. 4.

To give a conveniently tightened version of this argument:

(1) Mary knows everything there is to know about brain states and their properties.
(2) It is not the case that Mary knows everything there is to know about sensations and their properties.
Therefore, by Leibniz's law,
(3) Sensations and their properties ≠ brain states and their properties.

It is tempting to insist that we here confront just another instance of the intensional fallacy discussed earlier, but Jackson's defenders[12] insist that 'knows *about*' is a perfectly transparent, entirely extensional context. Let us suppose that it is. We can, I think, find at least two other shortcomings in this sort of argument.

The First Shortcoming: This defect is simplicity itself. 'Knows about' may be transparent in both premises, but it is not *univocal* in both premises. (David Lewis[13] and Laurence Nemirow[14] have both raised this same objection, though their analysis of the ambiguity at issue differs from mine.) Jackson's argument is valid only if 'knows about' is univocal in both premises. But the kind of knowledge addressed in premise 1 seems pretty clearly to be different from the kind of knowledge addressed in (2). Knowledge in (1) seems to be a matter of having mastered a set of sentences or propositions, the kind one finds written in neuroscience texts, whereas knowledge in (2) seems to be a matter of having a representation of redness in some prelinguistic or sublinguistic medium of representation for sensory variables, or to be a matter of being able to *make* certain sensory discriminations, or something along these lines.

Lewis and Nemirow plump for the "ability" analysis of the relevant sense of 'knows about', but they need not be so narrowly committed, and the complaint of equivocation need not be so narrowly based. As my alternative gloss illustrates, other analyses of

---

[12] See, for example, Keith Campbell, "Abstract Particulars and the Philosophy of Mind," *Australasian Journal of Philosophy,* LXI, 2 (June 1983): 129–141.

[13] "Postscript to 'Mad Pain and Martian Pain'," *Philosophical Papers,* vol. I (New York: Oxford, 1983).

[14] Review of Thomas Nagel, *Mortal Questions, Philosophical Review,* LXXXIX, 3 (July 1980): 473–477.

'knowledge by acquaintance' are possible, and the charge of equivocation will be sustained so long as the type of knowledge invoked in premise 1 is distinct from the type invoked in premise 2. Importantly, they do seem very different, even in advance of a settled analysis of the latter.

In short, the difference between a person who knows all about the visual cortex but has never enjoyed a sensation of red, and a person who knows no neuroscience but knows well the sensation of red, may reside not in *what* is respectively known by each (brain states by the former, qualia by the latter), but rather in the different *type* of knowledge each has *of exactly the same thing*. The difference is in the manner of the knowing, not in the nature of the thing(s) known. If one replaces the ambiguous occurrences of 'knows about' in Jackson's argument with the two different expansions suggested above, the resulting argument is a clear non sequitur.

(a) Mary has mastered the complete set of true propositions about people's brain states.
(b) Mary does *not* have a representation of redness in her prelinguistic medium of representation for sensory variables.
Therefore, by Leibniz's law,
(c) The redness sensation ≠ any brain state.

Premises a and b are compossible, even on a materialist view. But they do not entail (c).

In sum, there are pretty clearly more ways of "having knowledge" than having mastered a set of sentences. And nothing in materialism precludes this. The materialist can freely admit that one has "knowledge" of one's sensations in a way that is independent of the scientific theories one has learned. This does not mean that sensations are beyond the reach of physical science. *It just means that the brain uses more modes and media of representation than the simple storage of sentences.* And this proposition is pretty obviously true: almost certainly the brain uses a considerable variety of modes and media of representation, perhaps hundreds of them. Jackson's argument, and Nagel's, exploit this variety illegitimately: both arguments equivocate on 'knows about'.

This criticism is supported by the observation that, if Jackson's form of argument were sound, it would prove far too much. Suppose that Jackson were arguing, not against materialism, but against dualism: against the view that there exists a nonmaterial substance—call it "ectoplasm"—whose hidden constitution and nomic intricacies ground all mental phenomena. Let our cloistered

Mary be an "ectoplasmologist" this time, and let her know$_1$ everything there is to know about the ectoplasmic processes underlying vision. There would still be something she did not know$_2$: what it is like to see red. Dualism is therefore inadequate to account for all mental phenomena!

This argument is as plausible as Jackson's, and for the same reason: it exploits the same equivocation. But the truth is, such arguments show nothing, one way or the other, about how mental phenomena might be accounted for.

The Second Shortcoming: There is a further shortcoming with Jackson's argument, one of profound importance for understanding one of the most exciting consequences to be expected from a successful neuroscientific account of mind. I draw your attention to the assumption that even a utopian knowledge of neuroscience *must* leave Mary hopelessly in the dark about the subjective qualitative nature of sensations not-yet-enjoyed. It is true, of course, that no sentence of the form "*x* is a sensation-of-red" will be deducible from premises restricted to the language of neuroscience. But this is no point against the reducibility of phenomenological properties. As we saw in section I, direct deducibility is an intolerably strong demand on reduction, and if this is all the objection comes to, then there is no objection worth addressing. What the defender of emergent qualia must have in mind here, I think, is the claim that Mary could not even *imagine* what the relevant experience would be like, despite her exhaustive neuroscientific knowledge, and hence must still be missing certain crucial information.

This claim, however, is simply false. Given the truth of premise 1, premise 2 seems plausible to Jackson, Nagel, and Robinson only because none of these philosophers has adequately considered how much one might know if, as premise 1 asserts, one knew *everything* there is to know about the physical brain and nervous system. In particular, none of these philosophers has even begun to consider the changes in our introspective apprehension of our internal states that could follow upon a wholesale revision in our conceptual framework for our internal states.

The fact is, we can indeed imagine how neuroscientific information would give Mary detailed information about the qualia of various sensations. Recall our earlier discussion of the transformation of perception through the systematic reconceptualization of the relevant perceptual domain. In particular, suppose that Mary has learned to conceptualize her inner life, even in introspection, in terms of the completed neuroscience we are to imagine. So she does not identify her visual sensations crudely as "a sensation-of-black", "a sensation-of-grey", or "a sensation-of-white"; rather she identifies them more revealingly as various spiking frequencies in the *n*th layer of the occipital cortex (or whatever). If Mary has the relevant neuroscientific concepts for the sensational states at issue (viz., sensations-of-*red*), but has never yet been *in* those states, she may well be able to imagine being in the relevant cortical state, and imagine it with substantial success, even in advance of receiving external stimuli that would actually produce it.

One test of her ability in this regard would be to give her a stimulus that would (finally) produce in her the relevant state (viz., a spiking frequency of 90 hz in the gamma network: a "sensation-of-red" to us), and see whether she can identify it correctly *on introspective grounds alone,* as "a spiking frequency of 90 hz: the kind a tomato would cause." It does not seem to me to be impossible that she should succeed in this, and do so regularly on similar tests for other states, conceptualized clearly by her, but not previously enjoyed.

This may seem to some an outlandish suggestion, but the following will show that it is not. Musical chords are auditory phenomena that the young and unpracticed ear hears as undivided wholes, discriminable one from another, but without elements or internal structure. A musical education changes this, and one comes to hear chords as groups of discriminable notes. If one is sufficiently practiced to have absolute pitch, one can even name the notes of an apprehended chord. And the reverese is also true: if a set of notes is specified verbally, a trained pianist or guitarist can identify the chord and recall its sound in auditory imagination. Moreover, a really skilled individual can construct, in auditory imagination, the sound of a chord he may never have heard before, and certainly does not remember. Specify for him a relatively unusual one—an F#9th*add*13th for example—and let him brood for a bit. Then play for him three or four chords, one of which is the target, and see whether he can pick it out as the sound that meets the description. Skilled musicians can do this. Why is a similar skill beyond all possibility for Mary?

"Ah," it is tempting to reply, "musicians can do this only because chords are audibly structured sets of elements. Sensations of color are not."

But neither did chords seem, initially, to be structured sets of elements. They also seemed to be undifferentiated wholes. Why should it be unthinkable that sensations of color possess a comparable internal structure, unnoticed so far, but awaiting our determined and informed inspection? Jackson's argument, to be successful, must rule this possibility out, and it is difficult to see how he can do this *a priori*. Especially since there has recently emerged excellent empirical evidence to suggest that *our sensations of color are indeed structured sets of elements.*

The retinex theory of color vision recently proposed by Edwin Land[15] represents any color apprehendable by the human visual system as being uniquely specified by its joint position along three vertices—its reflectance efficiences at three critical wavelengths, those wavelengths to which the retina's triune cone system is selectively responsive. Since colors are apprehended by us, it is a good hypothesis that those three parameters are represented in our visual systems and that our sensations of color are in some direct way determined by them. Sensations of color may turn out literally to *be* three-element chords in some neural medium! In the face of all this, I do not see why it is even briefly plausible to insist that it is utterly impossible for a conceptually sophisticated Mary accurately to imagine, and subsequently to pick out, color sensations she has not previously enjoyed. We can already foresee how it might actually be done.

The preceding argument does not collapse the distinction (between knowledge-by-description and knowledge-by-acquaintance) urged earlier in the discussion of equivocation. But it does show that the "taxonomies" that reside in our prelinguistic media of representation can be profoundly shaped by the taxonomies that reside in the lingustic medium, especially if one has had long practice at the observational discrimination of items that answer to those linguistically embodied categories. This is just a further illustration of the plasticity of human perception.

I do not mean to suggest, of course, that there will be no limits to what Mary can imagine. Her brain is finite, and its specific anatomy will have specific limitations. For example, if a bat's brain includes computational machinery that the human brain simply lacks (which seems likely), then the subjective character of *some* of the bat's internal states may well be beyond human imagination. Clearly, however, the elusiveness of the bat's inner life here stems not from the metaphysical "emergence" of its internal qualia, but only from the finite capacities of our idiosyncratically human brains. Within those sheerly structural limitations, our imaginations may soar far beyond what Jackson, Nagel, and Robinson suspect, if we possess a neuroscientific conceptual framework that is at last adequate to the intricate phenomena at issue.

---

[15] "The Retinex Theory of Color Vision," *Scientific American* (December 1977): 108–128.

I suggest then, that those of us who prize the flux and content of our subjective phenomenological experience need not view the advance of materialistic neuroscience with fear and foreboding. Quite the contrary. The genuine arrival of a materialist kinematics and dynamics for psychological states and cognitive processes will constitute not a gloom in which our inner life is suppressed or eclipsed, but rather a dawning, in which its marvelous intricacies are finally *revealed*—most notably, if we apply ourselves, in direct self-conscious introspection.

PAUL M. CHURCHLAND

University of California/San Diego

BOOK REVIEWS

Perception: A Representative Theory. FRANK JACKSON. New York: Cambridge University Press, 1977. x, 180 p. $27.50.

Frank Jackson's admirable book forcefully argues that all visual experiences involve seeing sense data, that sense data are mental objects, and that seeing a material object consists in seeing a sense datum that, in a sense Jackson defines, "belongs to" that material object. These claims constitute, according to Jackson, the core of a representative theory of visual perception. Jackson argues in the course of chapter III that, like visual sense data, bodily sensations are mental objects; otherwise he confines his attention to visual perception. But his arguments would very likely adapt successfully to other sensory modalities.

Jackson's conclusions will strike many readers as somewhat extravagant. Nonetheless, his arguments for these conclusions are invariably models of clarity, conciseness, and cogency. He builds his case for sense data and representationalism with meticulous care, and helpfully directs attention to the connections among the various steps in his over-all argument. Accordingly, I first describe the structure of that argument. I then turn to several steps that seem especially open to objection. Finally, I sketch a way one might enjoy the advantages of Jackson's account while avoiding its more extravagant aspects.

0022-362X/85/8201/0028$01.40

I

In chapter I Jackson sets out to show that, whenever any object of visual perception exists, there exists an immediate object, and that these immediate objects are always colored shapes. An object of perception is *immediate,* roughly, if we cannot analyze what it is to perceive that object in terms of perceiving some other object. This definition guarantees that no object of perception can exist unless some immediate object exists. So all Jackson must show is that, in visual perception, these immediate objects are always colored shapes (I.7/8).*

Jackson begins by distinguishing cases in which it is obvious that some object of perception exists from cases in which it is not. Veridical and illusory perception plainly do involve such objects; whether visual hallucinations, such as afterimages, also do is controversial. Jackson then argues that all theories of perception agree that the immediate objects of veridical perception, defined as above, are colored shapes. For representationalism, direct realism, and idealism all insist on this, differing only "about the ontological status of the coloured expanse" (21), specifically, about whether it is a physical surface or something mental. Jackson concludes that we are warranted in claiming that the immediate objects in these two cases are colored shapes. By contrast, theories differ about whether any object of perception exists in cases of visual hallucination. But here Jackson plausibly argues that, if hallucinating does involve seeing some object, that object would have to be a colored shape. Moreover, there could be no object of visual hallucination that was not an immediate object (I.9). Veridical, illusory, and hallucinatory perception plainly exhaust the cases. Jackson concludes that, whenever some object of visual perception exists, the immediate object is a colored shape.

One might challenge Jackson's assumption that representationalism, direct realism, and idealism cover the available range of serious theories. But even if they do, Jackson's contention that all three theories imply that immediate objects are colored shapes is itself problematic. Direct realists use 'colored shape' to refer to physical surfaces, whereas representationalists and idealists apply the term instead to something mental. It therefore is questionable whether these theorists mean the same thing by 'colored shape'. If not, their superficial agreement here relies on an equivocation. To avoid this conclusion, Jackson must insist that words for color and for spatial properties mean the same thing, whether applied to mental or physical objects. Later (III.18–20) Jackson explicitly articulates and defends this "univocality thesis" (74), which plays a pivotal role throughout the book.

---

* Such references are to sections of chapters: here, to ch. I, secs. 7 and 8; other parenthetical references are to pages of Jackson's book.

After chapter I, only two things remain to be shown before Jackson can conclude that all visual experiences involve seeing sense data. Again, the argument proceeds by distinguishing veridical and illusory vision from visual hallucination. Sense data, presumably because they are given to the senses, are traditionally taken to have just the properties they appear to have. Chapter I does show that both veridical and illusory visual perception have immediate objects that are colored shapes, but not also that those immediate objects have the colors and shapes they seem to have. Chapters II and IV argue this. By contrast, chapter I does show that if immediate objects of visual hallucination ever exist, those objects would have to have their apparent colors, but not that such objects actually exist. Chapter III attempts to show they do.

Jackson's fundamental strategy, in both cases, is to consider certain indisputably true statements and to argue that, if taken at face value, these statements imply the claim he seeks to establish. He then argues that no analysis of these statements is possible that allows us to avoid that result. His arguments for and against particular analyses are generally probing and illuminating, even when the over-all argument is not completely convincing.

Jackson's argument for objects of hallucinatory vision relies on statements about afterimages. Such statements ostensibly treat afterimages as objects; they purport to individuate afterimages, and attribute properties to them. Afterimages plainly are not physical objects. So, if they are objects at all, they must be mental objects (III.2). Jackson considers three ways one might seek to circumvent this result. We might, first, represent phrases that putatively designate afterimages as actually parts of larger, syntactically atomic predicates (III.6). The second way seeks, instead, to avoid the conclusion by characterizing afterimages as mental states rather than mental objects (III.5–9). The third, by analyzing perceptual statements adverbially, again avoids reference to mental objects (III.11–17). As Jackson notes, it is convenient to see the third strategy as a variant of the second, in which adverbially characterized mental events serve in place of mental states (63).

Jackson insists that, to achieve any "significant ontological reduction" (59), the second strategy must not countenance more than one mental state per person. We can identify, for reductive purposes, a single state with the person who is in the state, but Leibniz's law blocks this move if we have two or more distinct states (III.7/8). Similarly, Jackson in effect assumes, for the same reason, that the third strategy does not involve a plurality of mental states (III.11). Given this constraint, he argues convincingly that all three strategies make it impossible to capture various uncontroversial semantic connections that obtain among statements about afterimages. His most compelling argument shows that, so constrained, these strategies cannot distinguish having an afterimage that is red and green from having two afterimages, one red and the other green.

To show that immediate objects of veridical and illusory vision have the properties they seem to have, Jackson relies on statements about how things look. Such statements purport to talk about the looks things have; Jackson argues that we cannot both avoid commitment to the looks of things and save the semantic phenomena. Thus we cannot treat 'looks *F*' as syntactically atomic and still explain the semantic relation between looking *F* and being *F* (IV.3). We cannot analyze how things look in terms of our having beliefs about them (II.5–7) or in terms of their looking like other objects (II.4); Jackson invokes familiar and convincing arguments against analyzing the "phenomenal" sense of 'looks' by appeal to such "comparative" or "epistemic" senses.[1] Moreover, we can compare and distinguish visual images in qualitative terms that are independent of physical objects (IV.5, 8). Something's looking a particular way therefore implies the existence of its look. And evidently that look has the properties it seems to have. So, like the immediate objects of hallucination, those of veridical and illusory vision really are as they appear.

Since hallucinations are definitionally nonphysical, hallucinatory sense data are mental objects. But more is needed to show that the sense data of veridical and illusory perception are also mental. To help show this, chapter V argues that physical objects lack color. Our only reason for believing otherwise, Jackson urges, is that they cause in us perceptual experiences of color. But physical objects cause such experiences because of properties they have other than colors, and they would do so whether or not they were actually colored. For they cause these experiences solely by virtue of their molecular structure. And we cannot identify colors with such properties, because afterimages have color but lack those properties (127). So the belief that physical objects are colored is both unwarranted and idle (V.3-5). But the sense data in all visual perception are colored shapes. These sense data cannot, accordingly, be physical objects. They must, even in veridical and illusory perception, be mental objects.

---

[1] See, e.g., Roderick M. Chisholm, *Perceiving: A Philosophical Study* (Ithaca, N.Y.: Cornell, 1957), ch. 4.

As chapter VI convincingly argues, this conclusion does not jeopardize our knowledge of nonmental reality. We can obtain such knowledge by an inference to the best explanation of our mental sense data (VI.4-7), and the knowledge so obtained is no less secure than it would be according to a direct realist theory of perception (VI.8). In chapter VII, finally, Jackson defines what it is for a sense datum to "belong to" a physical object: The physical object must, first, cause the having of the sense datum; and, second, the spatial properties of the sense datum—its size and shape, and its distance and direction from the perceiver—must vary as functions of the corresponding spatial properties of the physical object, and so vary because of how the two are casually connected. Accordingly, Jackson concludes, we can analyze seeing a physical object as seeing a sense datum that belongs to that physical object (101; cf. VII.11).

II

Jackson's account commits him to a strikingly unintuitive conclusion, which he somewhat surprisingly embraces. When I see a physical object, it has a particular look. That look is a mental object whose properties are just as they appear. So, if what I see looks to be fifty feet away, a mental object exists that actually is fifty feet away (102/3). Indeed, if the sizes and shapes of afterimages are the same kinds of properties as those of physical objects, it is hard to see how they could help but have the same kind of external spatial location as well.

This consequence fits nicely with Jackson's definition of what it is for a sense datum to belong to a physical object. The sense datum must vary with the physical object not only with respect to size and shape, but also in regard to distance and direction from the perceiver. Jackson urges that it is no more problematic to locate perceptual mental objects outside the body in physical space than to assign pains the bodily locations they appear to have, and he has defended so locating pains earlier in the book (III.20). Indeed, no special problem arises about mental objects' existing outside the perceiver's body, since pains can be felt in phantom limbs (103). And, though Jackson does not so argue, if pains really are where they seem to be, then, at least in phantom-limb cases, neural events cause pains to occur instantaneously at some distance. If so, neural events could presumably also cause sense data to occur instantaneously fifty feet away. But these considerations will convince few that perceptual images or other mental objects can exist outside the body.

Jackson argues that commitment to a plurality of mental states has no ontological advantage over countenancing mental objects. Since one is numerically distinct from one's states, "to have a sensation [e.g.,] will be, as on the act-object theory, to be related to something other than oneself" (60). So the choice between states and objects "appears to be verbal rather than ontological" (59). Here Jackson considers only the number of items countenanced, and not their nature. Describing an afterimage as a mental object suggests it has a fair measure of independence, substantially more than if it were just a state of the perceiver. Conceiving of afterimages as objects thus camouflages the bizarreness of Jackson's claim that afterimages exist at some distance from perceivers. By contrast, if afterimages are mental states, they would arguably have to exist wherever the perceiver does. If so, and if afterimages are mental states rather than objects, Jackson is wrong that when afterimages look to be outside the perceiver they actually are. Jackson has bypassed an important consequence of the choice between states and objects.

Jackson convincingly shows that an account that invokes states at all must invoke a plurality of states for each perceiver. But he nowhere considers construing this plurality as a single state, consisting in the sum of the constituent states. Indeed, his insistence that perceptual and bodily sensations have the same spatial properties as physical objects actually invites this move. Spatially contiguous physical objects, if causally related in suitable ways, will be parts of a larger physical object. So it is natural to take contiguous states of a single perceiver to be parts of a larger whole. And, if this larger whole is a unitary sensory state, we can according to Jackson identify it with the perceiver, thereby achieving a "significant ontological reduction" (59). Such unitary states are not mere reductive constructs, but have an intuitive basis in our conception of a sensory field. One's concurrent visual sensations are not independent objects, but together constitute the unified content of one's visual field.

The difficult problem of whether sensations of distinct modalities belong to the same sensory field is irrelevant here. For one thing, Jackson recognizes that my being warm and my being happy are each unitary states that in "a natural, if philosophically difficult, sense . . . are not things over and above and distinct from me." Our having distinct states does not here keep us from identifying them with the perceiver, since they are of different kinds (59).

Several unitary sensory states, each of a distinct modality, should similarly be nothing over and above the perceiver. But we could in any case simply identify all sensations that have the same location, whatever their modality, and rely on properties special to each modality to preserve necessary distinctions.

The composite character of those unitary states also occasions no difficulty. Being happy and being warm each consist of component states. One generally is happy about distinct things, and being warm implies that a preponderance of one's bodily parts are warm. Jackson seems to think some problem exists about mental states having parts at all (e.g., 62, 64/5), but he never specifies what it is. Intuitively, the contents of a sensory field are distinguishable parts of a unified whole. That unity permits ontological reduction; there being parts allows us to save the semantic phenomena.

Jackson's argument that physical objects lack color succeeds only if we cannot identify the colors of physical objects with their light-reflecting properties. Jackson insists that this identification fails, since afterimages have colors but cannot reflect light. Univocality is again pivotal for Jackson's argument, which evidently presupposes that the colors of afterimages and of physical objects are the same.

Jackson argues for univocality as the best explanation of the "striking fact" that words for colors and for spatial properties apply both to visual hallucinations and to physical objects (73). But he does not raise the possibility that these words might, in the two contexts, have distinct but systematically connected meanings. The only alternative he sees to univocality is that the terms in question are "radically ambiguous" (139). And then the dual use these terms have in many languages would be "a linguistic accident, a fantastic fluke" (73).

But these terms are neither univocal nor radically ambiguous. We use color words to classify physical objects. Physical objects presumably have properties that make this possible, and it is natural to take these properties to be the colors of the objects. But we also use color words to distinguish among visual sensations that, in standard circumstances, are caused by variously colored physical objects. This much hardly suggests that physical objects and visual sensations share a kind of property in virtue of which we classify both. The natural inference is rather that physical objects have properties in virtue of which we distinguish them, and that visual sensaitons have distinct properties that allow us to distinguish sensations that typically result from physical objects of various colors, thus making it possible for us to distinguish those physical objects.

We do not use 'red' to attribute the same property to a physical object and to a visual sensation; accordingly, 'red' means distinct things in the two contexts. But, because red sensations typically result from red physical objects, the two meanings are systematically connected. Parallel remarks apply to words for spatial properties.[2]

Notoriously, we cannot spell out precisely the standard circumstances in which, e.g., a red physical object looks red. Intertranslation is therefore impossible between statements about red physical objects and those about corresponding sensations (II.4, 76/7; also ch. II *passim*). So, if 'red' means two distinct things in those two contexts, we shall be unable to analyze one meaning in terms of the other. But Jackson maintains that, when terms have distinct but related meanings, we must always be able to analyze one meaning in terms of the other (74, 76/7, IV.3, 109–112). And, if 'red' cannot have distinct but related meanings and cannot be radically ambiguous, it must be univocal.

But it places excessive strain on the traditional analytical apparatus Jackson adopts at the outset (2/3) to suppose that we can always reveal analytic connections among semantically related terms. Analogical extensions of terms, in particular, resist such treatment. It is natural to think of terms such as 'burning' or 'dull' as having been extended by analogy to apply to bodily sensations. The same is credible about the application of color words to visual sensations.

If, *pace* Jackson, the colors of physical objects are distinct properties from those of visual experiences, pairs of properties for which we use the same terms will be intrinsically unrelated. The redness of visual sensations would thereby correspond to that of physical objects only contingently (cf. 139), as a result of the particular constitution of our perceptual apparatus. Jackson seems to assume, however, that meanings of terms always correspond to properties that the terms attribute (e.g., 73). If color terms ascribe unrelated properties to physical objects and to visual sensations, they would then have correspondingly unrelated meanings. Jackson's treatment of color and spatial terms avoids this consequence. Color terms apply only to sense data; thus no such term ascribes more than one property. And, since spatial terms ascribe the same properties to sense data as to physical objects, radical ambiguity is avoided there as well.

---

[2] Berkeley clearly held that at least some terms for sensible qualities fail to apply to things univocally. Thus, 'plane' and 'solid' apply primarily, on his view, to the immediate objects of touch, and only derivatively to the objects of sight. Berkeley sometimes seems to claim that such terms are radically ambiguous, as when he writes that the visual and tactile objects to which we apply these terms are "of a nature intirly different." But he also insists that planes and solids are both "equally suggested by the immediate objects of sight, [and] accordingly are themselves denominated plains and solids" [*A New Theory of Vision*, sec. 158, in *The Works of George Berkeley*, ed. A. A. Luce and T. E. Jessop (London: Thomas Nelson, 1948), vol. I; I am grateful to Robert Schwartz for this reference]. If so, the two applications of the terms, though based on distinct meanings, will nonetheless be systematically connected.

But connections between such pairs of properties need not hold singly, color by color and shape by shape; rather, one individual property can correspond to another by virtue of the position each has in two families of properties that are suitably connected as systematic wholes. Thus it is reasonable to regard the entire family of mental color properties, e.g., as corresponding, by virtue of relations among its members, to the family of physical colors. Intuitively, some such wholesale correspondence between mental and physical colors underwrites our systematic dual use of color terms. The relation between physical red and green, e.g., resembles the relation between the red and green of visual sensations. Parallel remarks again apply to words for spatial properties. Jackson seeks to achieve univocality by treating color and shape on different models. The present suggestion restores uniformity to our conceptions of color and shape by relinquishing an implausible univocality.

Jackson concedes that words like 'burning' do have distinct but related meanings as applied to physical objects and to bodily sensations (76). But he denies that the case of color words is similar. For "one does not know what is meant if someone says something is burning unless one knows whether the something is mental or physical." By contrast, "I know precisely what I mean by saying that I am seeing a bright yellow flash, even [if] I do not know whether the flash is a light flash or a hallucination" (76). But what I mean here presumably hinges on what I believe. If I think there is a physical flash of light, I may mean to say it is that which is yellow; if I am unsure, I will mean to describe only my visual sensation. If you do not know which I believe, you may well not know what I mean.

III

Seeing something, according to Jackson, involves the perceptual state of having a sense datum (167, 171), and having a sense datum is "be[ing] related to something other than oneself" (60). Jackson sensibly counts something as mental if it could not exist without sentient creatures' existing (3); if something is "in the mind," it is "incapable of independent existence" (80). Thus, although Jackson sees mental objects as distinct from, and often external to, sentient creatures, mental objects are also ontologicaly dependent on them. A satisfactory account of what it is for a sentient creature to have a mental object must incorporate both aspects of the relation between physical and mental objects, according to Jackson. It is unclear that any credible account can do so. Nor does Jackson attempt such an account. Careful as he is to define what it is for a sense datum to belong to a physical object, he nowhere even hints at what it is for a sentient creature to have a sense datum.

These considerations recommend an account of perception that countenances sensory states, no matter what their number, and not mental objects. Mental states are preferable not, as Jackson suggests (III.5, 11), because they permit ontological reduction, but rather because states are manifestly dependent upon sentient creatures. Mental events would do equally well, since the dependence of events on objects is widely recognized. An adverbial account that invoked events would work as well as an account based on states.

But an account that invokes mental states or events will cause serious difficulties for univocality. If terms univocally apply to physical objects and mental states, such states would, e.g., be red and triangular. And it seems clear that states and events cannot be red or triangular in the very same way that physical objects are. Physical objects have color, size and shape because they have surfaces. Describing afterimages as mental objects muted the oddness of Jackson's assigning them spatial locations outside the perceiver. Similarly, talking about mental objects in general conceals the intuitive strangeness of supposing that sensory items could have the kind of properties that physical objects have.

This point can be put more abstractly. Mental objects, whatever their nature may turn out to be, would doubtless differ in important ways from ordinary physical objects. But mental states will differ in at least one crucial additional respect. States are, roughly, ways objects are, not objects. Objects can bear relations to other objects; but an object cannot be in an object, in the way it can be in a state. We characterize objects in part by reference to the states they are in. States are, in effect, properties of objects. Mental states are thus the wrong type of item to share any nontrivial properties with physical objects. Mental objects, by contrast, do not diverge in this way from physical objects. Jackson might well concur here; when adverbs modify other adverbs, he notes, they stand for properties different from those they stand for when they qualify verbs or adjectives (66, 68).

Jackson might reply that, though we do describe mental objects, such as afterimages, by saying that they are red and triangular, we do not so describe mental states, such as visual sensations or experiences. Rather, we describe visual experiences as being of something red and triangular, or of a red triangle. This usage, he might insist, reveals that perceptual states are covert relations perceivers bear to mental objects, in this case, to a red and triangular mental object.[3]

But the difference between our descriptions of afterimages and those of other visual experiences is not due to the ontological status of afterimages; these descriptions do not differ because afterimages are mental objects, whereas other visual sensations are mental states. Rather, when one has an afterimage, it is clear that one is not seeing a physical object. So we describe the afterimage using terms that we would have applied to a physical object, if we were seeing one. By contrast, when we take our visual sensations to result from some corresponding physical object, we signal our conviction that a physical object is involved by describing our experience as being of something red and triangular. The 'of' in such locutions makes tacit reference not to mental objects, but to physical objects. Accordingly, to direct attention to the intrinsic quality of the experience, rather than the presumed character of the physical object, we revert to describing the experience as simply red and triangular.

Issues of level apart, it may seem moot whether an account invokes mental objects or mental states. Every mental object will correspond to the state of having that object, and every state can be redescribed as a relation to an object. And, in both cases, terms that characterize physical objects will also apply, whether univocally or not, to mental items of some sort. A nonverbal difference that separates such accounts may be hard to find.[4]

But if the mental items one's theory countenances are objects, one can represent them as having properties in common with physical objects. If the mental items are states or events, different properties must be involved. So, unless something other than univocality recommends mental objects, they will stand or fall with univocality itself. Indeed, the intuitive independence mental objects have from perceivers parallels the external locations Jackson assigns them. These locations are the specifically spatial consequence, by way of univocality, of the ontological independence of mental objects.

---

[3] Compare John R. Searle's insistence that "[r]edness is not part of my visual experience . . . ; the experience is *of* something red, but is not itself a red experience" [*Intentionality: An Essay in the Philosophy of Mind* (New York: Cambridge, 1983), p. 131].

[4] In his illuminating "Minds and Ideas in Berkeley" [*American Philosophical Quarterly*, VI, 3 (July 1969): 198–207] George Pitcher remarks on the difficulty of isolating a clear issue that divides act-object from adverbial theorists, whose accounts are cast in terms of mental states (p. 204). Pitcher proposes that we construe act-object theorists as allowing experiences to exist of which we are unaware, whereas adverbial theorists, he suggests, rule this out. Jackson does permit visual experiences of which we are unaware (24–26); but it is unclear why mental-state and adverbial theorists would not have the same reasons to do so as well.

Even Jackson seems at one point to see that univocality, and not the number of mental items, is what decides between mental states and objects. For in formulating his "univocality thesis" he writes that "[t]here is little philosophical bite to the bald assertion that there are mental objects," since that only amounts to asserting that a plurality of mental phenomena exist. "The bite comes with our claim that the mental objects really [i.e., univocally] are red, triangular, in the foot, or whatever" (73/4).

The univocality of color terms, as Jackson explains it, may not seem to amount to much. Though physical objects lack color, Jackson insists that statements like 'The tomato is red' are not on that account false. Rather, such statements attribute redness to sense data that "normally" belong to the physical object (128). Since 'red' applies only to sense data, intertranslation between statements about physical objects and sense data is not an issue. So 'red' is univocal by default. The property of sense data it stands for is a property physical objects cannot have (139). Jackson sees parallels with what holds when we speak of physical objects' causing things, or of sentences' being true (I.7/8). "[C]ausal links putatively between objects are analysable in terms of causal relations between events" (167) and, though sentences cannot literally be true (61), we can understand attributions of truth to them as indirect attributions of truth to propositions (16/7). Similarly, the bearers of color are sense data, and we can attribute color to physical objects only obliquely.

No such easy explanation of the univocality of spatial words, however, is available to Jackson. Color terms can apply obliquely to physical objects only because colored sense data belong to those objects. So Jackson must have some way, independent of colors, to specify this relation of belonging. To define that relation, Jackson needs a property that bridges the gap between sense data and physical objects. And only spatial terms are available. To explain belonging, Jackson must thus assume that physical objects and sense data both literally have the same spatial properties. Univocality here holds not by default, as with color terms, but because the univocality of color terms requires it.

Here too univocality is most troubling. Even if colors seem at home in the mind, literal spatial extension does not. Moreover, it is intuitively difficult to accept locating mental objects in space, at places also occupied by physical objects, and outside perceivers' bodies. Such locations are avoidable only by ad hoc adjustments on univocality. And, though Jackson disagrees (80), many would insist that an object's being in space along with physical objects implies that its presence is detectable in at least some of the ways physical objects are. These considerations doubtless explain why some who concur with Jackson about univocality go on to locate mental objects in a special "mysterious space" (103),[5] a move which Jackson rightly ridicules.[6]

But one need not have extended mental objects, whether in physical or "mysterious" space, to correlate physical objects with visual sensations in spatial terms. It will suffice if visual sensations simply have a range of properties that resemble and differ from one another in ways isomorphic to the resemblances and differences among the strictly spatial properties of physical objects. For then we can define a sensation as belonging to a particular physical object just in case those properties of the sensation which correspond to the physical spatial properties vary with the spatial properties of that physical object.

Univocality will then fail. For these properties of sensations are not literally spatial properties, but rather sensory analogues or correlates of spatial properties. It is presumably in virtue of those correlates that visual sensations fit together into a unified field and, conversely, that one can think of the visual field as consisting of parts. Parallel remarks apply to the other modalities.

We are now in a position to handle color. On Jackson's view, we call a physical object red, by courtesy, if red sense data "normally" (128) belong to it. On the present account, physical objects and visual sensations each have distinct properties in virtue of which we call them red, and the physical objects we count as red normally cause the visual sensations we call red.

The tendency to see the colors and shapes of perceptible physical objects as being the same as those of perceptual mental states is no more than the error, which Descartes warned against, of supposing that nonmental objects are "similar to the ideas which [they] caused." Descartes held that this error stems from our ignorance about the nature of bodily objects.[7] By contrast, the contemporary tendency to assimilate the two kinds of properties seems to result from our relative ignorance about the nature of mental states. We have such scant knowledge about mental states that we are tempted to model the properties of mental states on those of physical objects.

---

[5] See, e.g., Bertrand Russell's distinction between public and private space in *The Problems of Philosophy* (London: Oxford, 1912), pp. 29–33, and elsewhere.

[6] There exists some temptation to think that physical spatial relations obtain among the contents of our visual field. If this were so, those contents would presumably have spatial locations and, hence, other physical spatial properties, in just the way that physical objects do. But no temptation also exists to try to locate the entire field this way. Indeed, so locating it seems impossible. This apparent impossibility seems to demand an account based on analogues of spatial properties of physical objects, rather than those properties themselves.

Jackson might complain here that the mental correlates of the present account stand in need of explanation. Thus his main objection about a related account[8] is that it fails to provide "an alternative theory" (112), but only "draw[s] attention to the need for a [theory]" (111). By 'theory' Jackson here means an analysis of perceptual statements. But his analysis leaves at least as much unexplained as does the current account. Without knowing what it is for sentient beings to have sense data, e.g., it cannot be at all clear how physical objects cause sense data and, hence, why sense data belong to physical objects. And the present view faces no special difficulty about how we infer from mental properties to their physical counterparts. For whether the mental properties are mere correlates of or the same as the physical counterparts, we infer, as Jackson argues, to the best explanation of our sensory experiences. Since these issues do not favor either account over the other, Jackson's view has only univocality to recommend it. And, without univocality, we can avoid the perplexing extravagances of Jackson's ontology.

Despite the foregoing objections, *Perception* is a splendid book. It is rich in careful, illuminating argument, and abounds in detailed discussions that are subtle, precise, and penetrating. For example, the arguments for Jackson's analysis of seeing that something is so in terms of seeing objects (VII.4–7) are masterful. This book is certain to remain important for anybody interested in conceptual treatments of perception and mind.

DAVID M. ROSENTHAL

City University of New York
Graduate School and Lehman College

---

[7] Meditation VI, *The Philosophical Works of Descartes*, vol. I, Elizabeth S. Haldane and G. R. T. Ross, tr. (New York: Cambridge, 1967), p. 188.

[8] Derived from G. E. M. Anscombe's "The Intentionality of Sensation," in R. J. Butler, ed., *Analytical Philosophy*, second series (Oxford: Basil Blackwell, 1965): 158–181.

Philosophical Papers, Volume I. DAVID LEWIS. New York: Oxford University Press, 1983. 285 p. Cloth $29.95, paper $10.95

This valuable volume collects David Lewis's main publications from 1966 to 1980 in the areas of ontology, philosophy of mind, and philosophy of language. To eight of the papers are appended new Postscripts containing elaborations, retractions, and response to criticism. Many of these papers are classics of contemporary philosophy. All are written in his pellucid, unpompous style, and with utter honesty in dealing with objections. In his Introduction, Lewis notes that certain themes recur throughout the essays: among them are modal realism, materialism, and a qualified functionalism. I will comment on two of these themes, and add some remarks about his account of attributions of first-person attitudes.

Lewis holds that possible worlds are enormous concrete particulars. The actual world is just one out of many such particulars: for the nonactual worlds and their inhabitants do not, Lewis says, differ in kind from the actual world and its inhabitants. Modal operators are to be analyzed as quantifiers over these worlds; the word 'actual' is an indexical, analogous to the spatial indexical 'here'. Lewis's strongest opponent on this topic is a moderate modalist: the moderate modalist takes the modal operators as primitive, but countenances, for special purposes, constructed possible worlds built up by principles stated using the primitive operators from actual objects and properties. No very strong argument against the moderate modalist emerges from these essays.

Lewis says that only the extreme modal realist offers an instructive analysis of necessity (158). The instructiveness presumably lies in the fact that, on his theory, all instances of such schemata as '$\Box p \supset p$' and '$\Box p \supset \Diamond p$' when fully analyzed reduce to instances of quantificationally valid schemata, and so their truth is explained. But the extreme modal realist cannot offer this explanation for the truth of the quantificational principles themselves and must explain their validity some other way. Whatever the extreme modal realist says on the second pass about the quantificational principles, the moderate modalist can say on the first pass about the modal principles.

Two other unretracted arguments are broached in these papers in connection with modal realism. One proceeds from the expressive power of a first-order language quantifying over nonactual worlds and individuals, a power not present in a language with modal op-

0022-362X/85/8201/0042$00.50

erators and "actualist" quantifiers. One way to invoke this point is to argue that modal locutions of English in fact have a sense uncapturable in the operator language, and must be analyzed as quantifiers over *possibilia*. "There could have been things which do not, in fact, exist" is one example. But it has become well known that the operator language can express this if enriched either with an "actually" operator or with a propositional quantifier[1]; a truth-theoretic semantics for these additional devices need not quantify over *possibilia*. Alternatively, the point may be invoked by appeal to what may seem explicit quantification over *possibilia* in English: "In every circumstance in which . . . ," "Different possible people will result from these two policies." But the need for first-order regimentations here is one thing, the nature of the objects so quantified over is another. The moderate modalist as described allowed for such sentences, provided that their quantifiers can be taken as ranging over objects constructed from actual entities. He will require further argument for the conclusion that the nonactuals apparently quantified over are of the same kind as concrete actual individuals.

The second argument broached is that the indexical analysis of actuality explains why skepticism about "I am in the actual world" is absurd (19); and could Lewis not add that genuine indexicality requires some form of modal realism? The moderate modal realist, though, may say, as indeed Robert Merrihew Adams did say,[2] this: propositions are neo-Russellian entities whose constituents include objects and properties, and possible worlds are (at least) compossible sets of such propositions. To say that one of these worlds is actual is to say that all propositions in it are true *tout court*. So to doubt that one is in the actual world is to doubt that it is true that one exists—a skepticism even Descartes overcame.

These points may make it sound as if, apart from the extreme realist's talk of "being of the same kind," there is little difference in what he and his moderate opponent allow as intelligible. But the differences between them have other consequences. The moderate modalist's conception is, for instance, intrinsically hierarchical in a way his opponent's need not be. Suppose the modalist regards worlds as sets of neo-Russellian propositions, and takes these propositions to be represented by set-theoretic constructions from their constituents. Then no world can contain propositions about itself: the world can be built up only from objects available before it is built up—otherwise there would ultimately be some violation of the Axiom of Regularity. The moderate modalist's notion of a world inherits all the complications and all the virtues of an indefinitely extensible concept.

---

[1] See, *inter alia*, Kit Fine's "Postscript" in *Worlds, Times and Selves* (London: Duckworth, 1977).

[2] "Theories of Actuality," *Noûs*, VIII, 3 (September 1974): 211–231, at p. 225 ff.

I now turn to Lewis's functionalism. For Lewis, our concept of pain, like that of any other mental state, is the concept of a state apt to stand in certain causal relations to stimuli, to other mental states, and to behavior. He allows that another member of our natural kind can be in pain without being in a state that has this role, provided he is in the state *S* which occupies that causal role for us. Conversely, Martians can be in pain without being in state *S* provided they are in the state that occupies the causal role of pain for them. Lewis denies, as in consistency he has to, the possibility of a being who is in pain but who is not in state *S*, is not in the state that occupies the proper causal role of pain for him, and is unique in respect of his kind. To the objection that his account mistakenly omits what it is like to be in pain, Lewis answers that anything that feels like pain is pain; indeed "it is an impossible mistake to make" (130). As for knowing what it is like to be in pain, to have such knowledge is not to possess a special sort of information; it is rather, he says, to possess certain abilities—"abilities to recognize, abilities to imagine, abilities to predict one's behavior by imaginative experiments" (131).

Lewis's remark that the mistake is impossible seems comparable to the remark that it is impossible to make the mistake of offering a noncausal theory of memory, on the ground that memory is necessarily a causal notion. So it is, but it does not follow that a theorist's account entails that it is. If Lewis's remarks are not to beg the question, he has to do two things. First he has to give a functional characterization of the distinction between those mental states such that there is something it is like, subjectively, to be in them and those mental states of which that is not so; and second he has to show that from his favored account of the proper causal role of pain it follows that any subject counted as being in pain by his criteria falls on the right side of this distinction. I myself have no idea of whether the first task can be carried out, but believe that, if it can, the resulting theory will have to give much greater consideration to the notion of attention than has been common in the philosophical literature.

Lewis's additional remarks on *knowing* what it's like will not answer the doubts of one who thinks that the first task cannot be carried out. The problem is that the ability account—which in itself that doubter can accept—characterizes the abilities in relation to such states as imagination and experience (recognition). These are states such that there is something it is like to be in them: if they are not functionally explicable, knowing what it's like will not have been functionally explained either.

Finally, and even more briefly, I consider Lewis on first-person, *de se*, attitudes. These he takes to be relations between a person (slice) and a property. So, intuitively, in having the *de se* belief that he is happy, a subject locates himself among those possible objects which are happy; self-location by time slices captures present-tense beliefs. Lewis emphasizes the irreducibility of *de se* to *de dicto* attitudes and the role of *de se* attitudes in deliberation, action, and perception.

This apparatus has difficulty in properly accommodating perceptual demonstratives, an equally distinctive category of thought constituent. Can we identify a perceptual demonstrative 'that cat' with the complex mixed descriptive-demonstrative 'the cat causing in the way required for perception such-and-such features of my current sensory experience'? If so, Lewis has already captured the category. But there are two objections. One is that it is implausible that a subject has to possess the concept of perception and of sense experience in order to be ascribed attitudes with this relatively primitive level of content. More important, thoughts involving the perceptual demonstrative are thoughts a subject cannot have unless at the time he *has* a sensory experience that helps to determine the referent (if any) of the perceptual demonstrative; whereas enjoying a sensory experience is not required to have thoughts involving the complex descriptive-demonstrative constituent available to Lewis. A corresponding point could be made for memory images vis-à-vis memory demonstratives, and, in the case of sophisticated thinkers, for demonstrative thought about the subject's own experiences themselves. Lewis might introduce an array of different ways in which a subject can be related to a property, one for self-location, others for different sorts of perceptual demonstratives, and so forth. It may well be then—without detriment to either—that such a modified Lewisian theory is intertranslatable with a neo-Fregean theory that quantifies over irreducibly demonstrative modes of presentation: for the different ways of being related to a property seem to be in one-one correspondence with the types of demonstrative modes of presentation such a neo-Fregean recognizes.

CHRISTOPHER PEACOCKE

Center for Advanced Study in the Behavioral Sciences, Stanford
New College, Oxford

The Miracle of Theism. J. L. MACKIE. New York: Oxford University Press, 1983. 268 p. Cloth $27.95, paper $9.95.

Faith and Rationality. ALVIN PLANTINGA and NICHOLAS WOLTERSTORFF, eds. Notre Dame: Notre Dame Press, 1984. 336 p. Cloth $22.95, paper $11.95.

Faith and Reason. ANTHONY KENNY. New York: Columbia University Press, 1983. 94 p. $17.50.

God and Skepticism. TERENCE PENELHUM. Boston: D. Reidel, 1983. xiii, 186 p. Dfl. 80.

The Miracle of Theism is an extremely clear, rigorous, fair-minded, and on the whole comprehensive treatment of arguments for and against the existence of God and of the rationality of belief in God. It is also, sadly, John Mackie's last book; he died in December 1981, three months after delivering the completed typescript to the publishers. He made clear and well-argued contributions over a wide area of philosophy, and his last work shows no decline in his critical powers. Like the present reviewer, he does not think that there are any deductively valid arguments from premises evident to most rational beings either to the existence of God or to the nonexistence of God. The real issue is whether there is enough inductive force in those arguments from such premises which can be represented as inductive arguments, taken together, to render the conclusion that there is a God more probable than not. Mackie's structure of argument is thus very similar to that of my own *The Existence of God,* against the detailed arguments of which a number of his arguments are explicitly directed. He finds the individual arguments for the existence of God which can be given inductive form—the cosmological argument, the argument from design, the argument from consciousness, the argument from miracles, the argument from religious experience, and the moral argument—to have very little inductive force; and the argument from evil to have such great inductive force against the existence of God, that he concludes that it is most improbable that there is a God.

Some of the detailed arguments as to the inductive strengths of traditional arguments turn on detailed but very important points of inductive logic which there is no space to discuss in a general review.[1] But it is worth mentioning one crucial point to which Mackie returns several times. The temporal orderliness of nature, its regular conformity to scientific laws, is, Mackie agrees, scientifi-

---

[1] I have replied to Mackie's detailed criticisms of myself in a brief paper "Mackie, Induction, and God", *Religious Studies,* XIX (1983): 385–391.

0022-362X/85/8201/0046$00.80

cally inexplicable. We can explain the operation of lower-level laws by the operation of higher-level laws, but we have to come to a stop somewhere. Either this orderliness is a brute fact or it is to be explained in some nonscientific way. Now, the advocate of a design argument claims, there is available to us a pattern of explanation, different from the scientific one: personal explanation. We regard as a possible pattern of explanation of some phenomenon (e.g., an arm movement or a book being left on the table) that it was brought about intentionally by some person for some purpose. That they were brought about intentionally by such and such a person does, if true, provide an explanation of phenomena. And the orderliness of the world (e.g., the fact that each electron over infinite time and space behaves in exactly the same way, i.e., has the same powers and liabilities) is something that cries out for explanation—explanation should stop with some singular source of order, not with vast coincidences in the powers and liabilities of objects. Explanation would be provided if there were a super-person (God) who brings about the movements of electrons in the way that a singer intentionally produces tuneful notes or a dancer regular movements of his body. Hence, the design argument claims, orderliness of nature is evidence for the existence of God.

Against all this, Mackie claims that the analogy to human intentional action is not available. In humans the connection between intention and its fulfillment may seem "very natural and direct". But (129) this directness is illusory. "There is in fact a complex causal chain joining whatever brain event is the correlate of my decision through nerves and muscles to the movement of my arm." The theist cannot postulate a similar complex path between God and the laws of nature, and so, Mackie claims, human intentional action provides no analogy for the supposed divine action. There is, however, a quick way out of this alleged difficulty for the design argument. We can say that the analogy is not with an intention to move an arm causing the arm movement, but with an intention to move an arm causing whatever brain state causes the arm movement; and more generally, with an intention to make a bodily movement causing whatever brain state causes that bodily movement. There is here a "natural and direct" pattern of mind/body causation, the design argument may claim, as an analogy available to those seeking an explanation for the regular behavior of physical objects.

For this and many similar reasons of detail, this reviewer holds that the balance of argument goes the other way from the way Mackie claims. Mackie does not make any substantial original contribution to the subject, but his book will make a good course book for many a course on the philosophy of religion. It covers a wide area and discusses in a very fair way many writers, classical and modern, of different theological persuasions.

By contrast, *Faith and Rationality* illustrates at length an important new approach to the philosophy of religion. It is one *The Miracle of Theism* does not discuss, but one to which both the other books reviewed are in part or whole a reaction. *Faith and Rationality* consists of a number of essays by like-minded writers centered on Calvin College, developing a "Reformed epistemology" which, in loose summary, is the view that religious belief does not in general require to be justified by arguments, but may be held with perfect rationality without such backing. Or rather, although most of the contributors to this volume are in the above sense Reformed epistemologists, the most they purport to show by arguments is that the arguments of others to show Reformed epistemology mistaken do not work. When the theses for which they argue are spelled out, they contain many negatives. The opening 77-page paper by Alvin Plantinga sets the tone for the rest of the volume. His paper contains the most developed version of the thesis for which he has argued in other papers over the last few years (a thesis not discussed by Mackie). This is the thesis that *no*body known to Plantinga has shown that there are *no* circumstances in which some person or other may rationally treat "There is a God" as a basic belief. A basic belief or proposition is one that the believer believes not on the basis of reasons provided by some other belief. A rational or justified belief in Plantinga's understanding is one such that if a person believed it he would fail in *no* epistemic obligations; i.e., it is an epistemically permissible though not obligatory belief. A belief *B* is "properly basic" for a person *P* in circumstances *C* if *B* can justifiably be believed by *P* in *C* without his believing it on the basis of arguments (deductive or inductive) from other beliefs. Toward the end of the paper Plantinga suggests that the beliefs about which he is making his claim are not so much "There is a God" as such beliefs as "God is speaking to me" or "God has created all this", which deductively entail "There is a God" by immediate inference.

As my italics indicate, Plantinga's thesis is a pretty negative one. However, the only argument against the occasional proper basicality of theistic belief which Plantinga considers is that from classical foundationalism. Classical foundationalism holds that the only beliefs that are properly basic are those which are self-evident, evident to the senses, or incorrigible reports of experience; other beliefs are rational if and only if supported (deductively or inductively) by properly basic beliefs. Plantinga rightly argues that classical foundationalism is self-defeating; no one would be rational to believe it on its own understanding of rationality.

Although Plantinga rejects classical foundationalism, he does work within a foundationalist approach, in which some beliefs are properly basic and others are not and need to be supported by those which are. He disputes only the foundationalist's criterion of what is properly basic. But he does not even state, let alone attempt to argue for, a rival criterion of which propositions are properly basic. Such a criterion, he writes, can be established, if at all, only by induction from cases where it is clear that some proposition is taken as properly basic in certain circumstances, and cases where it is clear that some other proposition is not taken as properly basic in certain other circumstances. But he leaves open the possibility that theist and atheist may start from such different standard examples (including the crucial ones which are the subject of their dispute) that they cannot generalize to an agreed criterion. In that case one or other party to the dispute will remain in ignorance as to what are the criteria of proper basicality.

Plantinga notes that different kinds of proposition are properly basic in different circumstances. For example, I am justified in having "John is in pain" as a basic belief if John is displaying typical pain behavior and if various defeating conditions are not present (such as, e.g., that I am watching John act in a play). Plantinga then points out that there are various conditions that call forth belief in God—"guilt, gratitude, danger, a sense of God's presence, a sense that he speaks, perception of various parts of the universe" and he then claims that there are circumstances in which they do so justifiably. Now this is a much bolder positive thesis than the negative thesis that has been his central concern. But the positive thesis is simply stated, almost as an aside—no argument is given for it.

Similar well-argued highly cautious negative theses, with the occasional brief unargued bold positive claim come from other contributors to this volume. Nicholas Wolterstorff argues (with various qualifications) that a person is rational in holding a belief if he does *not* have adequate reason for ceasing to hold it (and, as with Plantinga, for Wolterstorff rationality is a matter of *not* violating epistemic obligations). And Wolterstorff then claims boldly that he sees *no* reason for believing that a person could *not* hold a belief in God as an immediate belief without having such reason. A slightly less cautious negative thesis comes in a very interesting and well-argued paper from William Alston. He argues that "the practice of objectifying certain ranges of experience in terms of Christian theology," e.g., describing some experience as God speaking to me, is such that one is justified in engaging in it in the sense that one does *not* have sufficient reasons for regarding it as *un*reliable, and that *no* stronger justificaiton can be given for our familiar secular way of objectifying some experience."

One trouble with all this concerns the Plantinga-Wolterstorff understanding of rationality, and the oddness of that understanding is drawn to our attention in a paper by George Mavrodes. On their understanding, a person's belief is rational if he is "within his intellectual rights" in holding it, if it would be "epistemically permissible" for him to hold it; and they seem to hold that often a given person in given circumstances would be rational in believing some proposition and equally rational in believing instead its negation. They assume that at that stage there is nothing more that reason can do to choose between the two. Yet in assessing beliefs as "rational" or "irrational" they seem to be concerned with rationality in a cognitive sense, and that is a matter of how much those beliefs are supported by what in a wide sense stares the subject in the face; that is, of how likely they are in some sense to be true, given the subject's experiences and present beliefs (including beliefs to which his attention is drawn in argument). Within this rough characterization of what the rational/irrational distinction is getting at, one can pick out different kinds of rationality which beliefs may possess: a belief may be rational insofar as it is likely to be true by objective criteria, by the subject's own criteria, in the subject's opinion, and so on. Yet for all such cognitive kinds of rationality, in general one proposition will be more likely to be true than its negation, and in that case for that kind of rationality the negation is not a rational belief. There may indeed be cases where both a proposition and its negation are equally likely to be true—more formally, both *p* and not-*p* have probabilities of $\frac{1}{2}$. But in that case there seems to me nothing at all rational in believing *p*—that would be going too far beyond reason. The rational man will believe, in line with the evidence, that *p* is as likely to be true as not. And if he believes that, he could not at the same time believe *p*. There may be justification for acting on the assumption of *p*, when you don't believe *p*; and for believing *p* when it is not likely to be true (e.g., that it helps toward salvation or gives peace of mind), but such justification does not make *p* a rational belief in the cognitive sense. If that is right, then the test of rationality will do more than knock out a few undesirable beliefs, leaving us with a vast pool to choose among; it will let in only desirable beliefs (and that will never allow both a proposition and its negation as equally admissible). Then the hurdle of rationality will be a harder one for a belief such as "There is a God" to surmount. If Plantinga and Wolterstorff are not concerned with rationality in this broad cognitive sense which I have characterized, they must explain what they are saying when they say that a belief is rational; and they must do this via some other cluster of notions than "intellectual rights", "epistemically permissible" and "epistemically obligatory". These notions are totally empty unless spelled out in less technical terms, or by examples.

The major problem, however, with the approach of *Faith and Rationality* is simply its negative character. What one would very much like to see from Plantinga is something more positive—some argument not merely that some one else's criteria of proper basicality are wrong, but that some criteria of his own are right—an argument that begins from a starting point that has some initial appeal to atheist and theist alike. For without such an argument his claim that belief in God is properly basic under certain circumstances will seem to the atheist (and to many a theist whose belief depends on argument) a false and unjustified claim. And even many a theist who takes his belief in God as basic, feels uncomfortable in the face of the classical foundationalist's challenge. He may agree with Plantinga that classical foundationalism is unsatisfactory as it stands and yet suspect that, subject to minor corrections, it is basically right—and so suspect that he is not justified in taking belief in God as basic. He would welcome a positive argument, starting from premises that seem to him initially more evident than the proposition that belief in God is properly basic for him, to show the latter proposition to be true (or at least to show true the proposition that belief in God is sometimes properly basic). The contributors to *Faith and Rationality* are sophisticated and able, and they certainly take the religious debate further. They would, I suspect, claim that they are only beginning their journey along a road which may yet yield something a lot more positive.

In *Faith and Reason* Anthony Kenny does attempt the very task which Plantinga has not yet faced, of stating a foundationalist criterion of proper basicality by which to assess the rationality of the belief that there is a God. He suggests that a belief is properly basic if and only if it self-evident or fundamental; evident to the senses or to memory; or defensible by argument, inquiry, or performance. A belief of some person is then rational if it is properly basic for that person, or supported by beliefs that are properly basic. By "fundamental" beliefs Kenny understands beliefs of the kind discussed by Wittgenstein in *On Certainty*, such as "The Earth has existed for many years past," "Cats do not grown on trees," and "Human beings often sleep"; they are fundamental, Kenny claims, in the sense that "they are basic in the noetic structure of every rational human being" and could not be given up without causing havoc to that structure, e.g., by casting doubt on the efficacy of human testimony in general.

With this wider understanding of rational belief, Kenny reaches the conclusion that "There is a God" could be properly basic for some person only if that belief were defensible by argument, i.e., a good deductive or inductive argument from other propositions rationally believed could be given for it, even if the believer in question did not hold it as a result of such an argument. So, in general, Kenny concludes, "There is a God" can be rationally believed, whether as a basic or non-basic belief, only if "natural theology is a possible discipline"; and on that crucial issue Kenny, unlike Mackie, remains agnostic.

Kenny's book is a very short one (94 small pages, grossly overpriced), and his position is outlined rather than defended at length. His criterion of proper basicality and its application to theism would be open to challenge by Plantinga along the lines which he stated to be appropriate for discovering the criteria of proper basicality. *Faith and Reason* was originally delivered as Bampton Lectures at Columbia University. It reads easily, and its arguments are clear; it must have been a pleasure to listen to. It does unfortunately move far too quickly over highly controversial matter, such as the epistemic worth of religious experience (which Alston treats in much greater depth in *Faith and Rationality*).

In *God and Skepticism* Terence Penelhum is concerned with the question of whether general philosophical skepticism is good for religion; whether the skeptic's claim that human reason can establish very little opens the way for faith, i.e., makes faith in some sense sensible or rational. Penelhum calls this doctrine, that skepticism is good for religion, Skeptical Fideism, and he detects two branches of it: conformist fideism and evangelical fideism. The conformist fideist claims that skepticism shows that it is sensible to follow the habits of ordinary life, without believing philosophical claims about a reality behind appearances; and so too in religion we may sensibly follow religious practice, without caring much about the cognitive credal claims of religion. The evangelical fideist claims that skepticism shows that in ordinary life it is not senseless to believe in a reality behind appearances in ordinary life, even though we can give not justification for our belief; and so it is not senseless in religion to believe that there is a God and to follow the demands of faith, even though there too we can give no justification for our belief. Penelhum's work is historically oriented, and he discusses the skepticism of the Greeks, especially Sextus Empiricus, the conformist fideism of Erasmus, Montaigne, and Bayle, and the evangelical fideism of Pascal and Kierkegaard. He then moves on to more modern writers, finding examples of conformist fideism in D. Z. Phillips and Don Cupitt (who add to more ancient conformist fideism, the view that worth-while religion does not or ought not have a cognitive credal content); and examples of a negative form of evangelical fidesim in Norman Malcolm and Plantinga.

Penelhum accepts the negative claim which he finds in Plantinga: that it is inconsistent for an atheist with many basic beliefs to accuse the theist of irrationality on the sole ground that he cannot provide arguments for his belief that there is a God. But he holds that skepticism of a kind that allows equal sense or equal rationality to following any of many rival ways or holding any of many incompatible beliefs is not a doctrine that a theist should welcome. For if the skeptic is right, atheism would also be a sensible or rational stance and is thus unlikely to be due to sin, as many a fideist claims that it is. The theist is in better shape if his beliefs are rational in a sense in which the atheist's beliefs are irrational; and in yet better shape if this can be shown by arguments from agreed premises that can be given to others, i.e., if natural theology is possible. This all seems very sensible.

Mackie, Kenny, and Penelhum all agree that religion needs natural theology. Mackie claims that it can't have it, Kenny says that he doesn't know whether it can or not, and Penelhum ventures no opinion on whether it can or not. Plantinga thinks that much religion can get on quite nicely without natural theology; but he does not say whether he thinks that the development of a successful natural theology would be some asset to religion.

Unless the Reformed epistemologist holds that everyone has or ought to have "There is a God" (or a proposition that immediately entails this) as a basic belief, he surely must hold that natural theology would be of use for converting those (few?) unfortunates who do not. True, he may hold that sin stands in the way of accepting good arguments; but he cannot hold that it prevents good arguments from having any effect at all. For if it did there would be no point at all in the enterprise of Reformed epistemology, as Mavrodes in effect points out. And natural theology will be of use not merely to those without "There is a God" as a basic belief, but also to those who have such a basic belief but are beginning to consider whether they are justified in holding it in their particular circumstances and do not find enough positive argument in Reformed epistemology to convince them. Even the Reformed epistemologist would find natural theology a useful ally. For those who do not have the required basic belief (and the more secular a society, the more of them there will be) and for those who hold that that belief can never be properly basic, natural theology is essential.

RICHARD SWINBURNE

Oriel College, Oxford

MEMO TO AUTHORS

The editors find that they can no longer afford to send to the printer manuscripts that are difficult for the compositor to read or footnotes that need to be completed in proof. They are therefore asking that manuscripts submitted be prepared according to the following guidelines:

1. All material, including the quotations and footnotes—especially footnotes—must be typed, double-spaced, on one side of the paper, with ample margins left and right, top and bottom.
2. No bibliographies will be published; bibliographical material should be put into the footnotes.
3. Footnotes should be complete, according to *Journal* style (see any issue). Journal citations should include name of journal in full, volume number, issue number, month or season, year, and pages. Book references should include publisher, city of publication, and date. Footnotes should be gathered at the end of the article.
4. Quotations should be checked by the author; capitalization, italics, spelling, etc., should match the original. Corrections should not have to be made (expensively) in proof.
5. If there are diagrams, they should be provided camera-ready by the author.
6. If special symbols are used, a list of those symbols should be provided with the manuscript.
7. Dot-matrix computer copy is extremely difficult to read. It is acceptable only if it is *all* double-spaced, if it is *not* justified, and if the print is dark enough so that punctuation marks, as well as letters, can be read.
8. Finally, photo copies on shiny paper are not acceptable. Ribbon copy is much preferred, but clear copies on bond paper are acceptable if necessary.

The editors are of course willing to read and evaluate manuscripts that do not meet these standards, provided they are genuinely legible and not unreasonably long. They will ask for revision of excessively long manuscripts, and they will require retyping of accepted manuscripts, if necessary, before they can be scheduled for publication.

The *Journal of Philosophy* is published monthly, in issues that (ordinarily) contain only 56 pages of copy; this limits the length of papers the editors can normally consider, to no more than 7,500 words, or eighteen JOURNAL pages. Manuscripts very much in excess of these limits will be returned unread.

The *Journal of Philosophy* does not publish papers that are primarily historical or expository; the editors urge that such manuscripts be submitted elsewhere.

JOHNSONIAN PRIZE IN PHILOSOPHY

The editors are pleased to announce that the winner of the Johnsonian Prize in Philosophy for 1984 is Robert J. Dunn, lecturer in philosophy at the University of Queensland, Australia, for his manuscript, *The Possibility of Weakness of Will.* His manuscript was sponsored by the University of Queensland. Previous Johnsonian prize winners have been: 1980, Daniel Bonevac, *Reduction in the Abstract Sciences* (Indianapolis: Hackett, 1981); 1981, Igal Kvart, *A Theory of Counterfactuals* (forthcoming); and 1982, Moreland Perkins, *Sensing the World* (1983). Honorable mention in 1981 went to Charles B. Guignon, *Heidegger and the Problem of Knowledge* (1983).

The Journal of Philosophy is pleased to announce that the Johnsonian Prize in Philosophy will be offered again in 1985. The prize consists of one thousand dollars and guaranteed publication of the prize-winning book. To encourage the work of unpublished philosophers, the editors of this JOURNAL plan to oversee the annual award of the Johnsonian Prize to an author who has previously published no book-length work and whose manuscript has been selected from among manuscripts sponsored by departments of philosophy.

Authors and departments who wish to submit manuscripts should observe the following guidelines:

1. Any graduate school of philosophy that confers the Ph.D. degree may submit one manuscript per year. It is not required that the work sponsored by a department be written by either a graduate or a faculty member of that department; it may, but need not be a revised version of a departmental dissertation. The closing date for this year's prize is August 1, 1985.
2. To qualify, the manuscript must be: (a) of book length (at least 40,000 words) and in proper form for publication as a book; (b) accompanied by an abstract of up to 2000 words; (c) written by someone who has not yet published any book-length work; (d) accompanied by two professional evaluations and a letter from the sponsoring department (the editors are depending on the departments to screen out all but the most important and original work); (e) uncommitted to a publisher.
3. Graduate schools who wish to participate should notify the JOURNAL in advance of submitting a manuscript for consideration, and, if possible, send evaluations and abstracts in advance.
4. Preference will be given to the work of younger philosophers, especially those who have not yet been granted tenure.

Announcement of the fifth award will be made in the January 1986 issue of the *Journal of Philosophy,* and $1,000 will be awarded to the prize-winning author. Some manuscripts may be given honorable mention.

Publication of the prize-winning book will be guaranteed by the Hackett Publishing Company after the author has made what revisions are deemed necessary by the editors, if possible within twelve months of the award. The sponsors hope that, if the schedule announced here is maintained, it will be possible to publish the fourth Johnsonian Prize winner in time for display at the APA Eastern Division meetings in 1986.

To Our Readers, concerning Copyright:

The editors and trustees of *The Journal of Philosophy* wish to draw to the attention of our readers the notice of permission for personal or educational copying which appears at the bottom of the masthead page of each issue of this JOURNAL.

In the last few years we have received many requests for permission to photocopy or mimeograph articles for classroom use; we have also received many requests from libraries for blanket permission to photocopy. And we are aware of course that a great deal of such photocopying has been done without permission. We are grateful to those of you who have asked for our permission.

The copyright law of 1976 grants certain new rights to libraries and educators who need to photocopy and certain new protections to copyright owners. In general, photocopies in excess of "fair use" copying cannot be made without authorization by the copyright owner. What constitutes "fair use" is not yet entirely clear, but some single-copying for personal use is permitted, and multiple-copying for classroom use is clearly limited: the article copied must be *short* (less than 2,500 words) or only an excerpt copied; the copying must be *spontaneous* (at the instance of the individual instructor, and needed so promptly as not to allow time to obtain permission); and *each copy must include a notice of copyright.*

The Association of American Publishers has been instrumental in setting up an organization, the Copyright Clearance Center, designed to make it as simple as possible for organizations to pay the fees required for copying *beyond fair use*. The *Journal* has joined the Center and thereby authorizes such copying under the terms established by the Center. The per-copy price for personal or educational copying beyond fair use appears on the first page of each article in the *Journal*. That price is 10¢/page, with a minimum of 50¢, unless some special arrangement is requested by the author. The *Journal* continues to give permission for multiple-copying for classroom use, but at 5¢/copy/page, where an instructor applies directly to us for our permission. We turn over to our authors all substantial sums ($15 or more) obtained from their work.

We wish also to point out that, in most cases, the cost of photocopying a *Journal* article is higher than the price of a *Journal* issue, even with our new monthly format ($2.00 for a current issue, $3.00 for a back issue of 64 pages). *Journal* issues are available from this office, singly as well as in volumes, back to 1904; prepaid orders are filled by return mail.

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXII, NO. 2, FEBRUARY 1985

PROBABILITIES AND CAUSES

THIS paper is about the connection between probabilities and causal explanation. I shall start by describing two familiar views on this matter. I shall call them the *standard view* and the *statistical-relevance view*. After that I shall introduce a different way of understanding the connection, which I shall call the *evidential view*. I shall conclude by arguing that once the evidential view is taken into account, the standard view and the statistical-relevance view lose their attractions.

I. THE STANDARD VIEW

On the standard view, the connection between probabilities and causal explanation is simple: one event explains another just in case the former gives the latter a high probability.[1] The higher the probability, the better the explanation. So the ideal case is the limiting one, where the explanans event determines the explanandum, and so allows us to deduce its occurrence; but one still has explanation, albeit of a weaker variety, when the probability falls short of unity, provided it is still reasonably high.

Thus, to take an example, if *all* people with streptococcal infections who take a certain dose of penicillin recover, one will be able to give a deductive explanation of why a particular sufferer who takes the dose recovers. But even if only 90% recover one can still explain, albeit slightly less satisfactorily, someone's getting better in terms of their taking the penicillin.

This standard view[2] faces a well-known difficulty. Suppose, as

---

[1] A terminological point. It will simplify the exposition if I can assume that all explanation of particular events is causal, that explaining one event by another is always a matter of showing the first to be caused by the second. As it happens, I think this assumption is right. But I shall not be concerned to defend it as such, so I would merely ask those who disagree to accept it as a terminological convenience, and remember that when I say 'explanation' I should be read as talking specifically about *causal* explanation.

[2] The "standard view" is of course the view developed by Carl G. Hempel. See his *Aspects of Scientific Explanation* (Glencoe, Ill.: The Free Press, 1965), ch. 12, pp. 376–411. However, it will simplify matters without, I hope, begging questions to work with an ideal type and avoid the details of Hempel's account.

0022-362X/85/8202/0057$01.80

above, that 90% of streptococcally infected people recover (R) upon taking penicillin (P): Prob(R/P) = 0.9. But suppose also that in a specific subclass of sufferers—those more than 70 years old, say—this proportion is radically lower. Suppose, for instance, that the proportion of recoverers among old (O) people who take the drug is only 10%: Prob($R/P\&O$) = 0.1.

And now consider the case of an old person who takes the penicillin and *does* recover. We've still got the probabilistic generalization applying to penicillin takers in general—Prob(R/P) = 0.9—and it covers this old person—he or she took penicillin (P). So, for all that's been said so far, it seems that we ought to be able to explain this particular recovery by the ingestion of the drug.

But surely we don't want this. Suppose that the same old person *hadn't* recovered, as was in fact very likely. Then we could have explained that too. For we have another generalization—Prob ($-R/P\&O$) = 0.9—and initial conditions—$P\&O$—in terms of which the *non*recovery could then have been given high probability.

What seems wrong here is the possibility of our explaining some outcome (the recovery) in the light of certain features of the actual situation, when we could have explained the opposite outcome (the nonrecovery) just as well if that had occurred, by mentioning other features of the situation.

But there is a natural enough way of ruling out this possibility. We can impose a requirement of *maximal specificity*: we can require that initial conditions mentioned in our explanation include *all* the relevant features of the situation. This then will stop the explanandum getting a high probability simply because certain relevant factors (the person is old) have been omitted from the explanans.

This maximal-specificity requirement is equivalent to the demand that the probabilities entering into our explanation should derive from a "homogeneous" partition of our reference class. Let me explain. In our example the original reference class was the class of all infected people. We then "partitioned" this reference class into the subclasses of those who took penicillin and those who didn't. But this partition failed to be "homogeneous" for recovery, in the sense that the probability of recovery depended not just on whether one took penicillin, but also on whether one was old. So we needed to partition again, into subclasses distinguished by age. A homogeneous partition for some outcome is a partition such that no further subdivisions will further alter the proportions with which that outcome is found within subclasses. Penicillin and age will give us a homogeneous partition for recovery if there are no further factors on which the probability of recovery depends.

The advantage of putting the maximal-specificity requirement in terms of homogeneity is that this makes explicit the precise sense in which the standard view should require an explanation to include all *relevant* factors. In our example, it wouldn't matter if the explanation omitted factors, like hair color, say, which were true of the person in question, but made no difference to the probability of recovery. What needs to be ensured is specifically that we include all those factors which *do* make such a probabilistic difference—that is, which are required to yield a homogeneous partition of the original reference class.

II. THE STATISTICAL-RELEVANCE VIEW

An alternative to the standard view is the view that the explanation of one event by another does not require that the former give the latter a *high* probability, but simply that it make it *more* probable than it would otherwise have been.[3]

On this view (the *statistical-relevance*[4] view henceforth), for someone's cigarette smoking (S) to explain their lung cancer (C) it is not required that smoking make cancer highly probable: Prob $(C/S) >> 0.5$, but only that it increase the probability: $\text{Prob}(C/S) > \text{Prob}(C/-S)$. It doesn't impugn the explanatory connection, on this view, that even among smokers the probability of cancer remains well below 50%: provided smoking increases the probability—from 1% to 10%, say—we have enough to allow explanations of cancer in terms of smoking.

As with the standard approach, there is an obvious (and indeed not unrelated) objection to the statistical-relevance approach. According to the statistical-relevance approach, the requirement for one event (A) to cause another (B) is that $\text{Prob}(B/A) > \text{Prob}(B/-A)$. Equivalently, that $\text{Prob}(B/A) > \text{Prob}(B)$; or, again, that $\text{Prob}(A\&B) > \text{Prob}(A)\text{Prob}(B)$. In short, the requirement is that A and B be "correlated," or "statistically associated," with each other. However, as generations of social science students know, "Correlation doesn't prove causation."

---

[3] This view is distilled from the writings of Patrick Suppes and Wesley Salmon. See Suppes, *A Probabilistic Theory of Causality* (Amsterdam: North-Holland, 1970), and Salmon *et al.*, *Statistical Explanation and Statistical Relevance* (Pittsburgh, Pa.: University Press, 1971). But again it will simplify matters to work with an ideal type. Indeed I should make clear that my "statistical-relevance view" differs from both Suppes's and Salmon's accounts on a number of points of substance.

[4] The term 'statistical-relevance' might suggest that the important difference between this approach and Hempel's is that it excludes irrelevant explanatory factors whereas Hempel fails to do so. But this is misleading. If Hempel's explicit formulations do sometimes leave room for irrelevant factors, this is a trivial matter, and easily remedied. One simply adds to the maximal-specificity requirement: include all relevant factors, the obvious corollary: omit irrelevant factors. The real difference between Hempel and the "statistical-relevance" alternative has nothing to with relevance as such; it is simply the difference between high probabilities and increased probabilities. Cf. sec. 5 of Salmon, "Statistical Explanation," in Salmon *et al.*, *op. cit.*

Thus, to take an example, although there is certainly a well-established correlation between smoking and cancer, this doesn't necessarily mean (as the cigarette companies wistfully point out) that smoking *causes* cancer. It may yet be that there is some hidden factor (a genetic predisposition, say) which, on the one hand, causes cancer, but which also, as an independent side effect, induces people to smoke. This would give us a perfectly good correlation between smoking and cancer, even though the first didn't cause the second.

To deal with this difficulty the statistical-relevance account needs to introduce a "no screener-off" requirement. Suppose we have an association between C (cancer) and S (smoking). That is, $\mathrm{Prob}(C/S) > \mathrm{Prob}(C)$. Then we shall say that some factor G (think of the genetic predisposition) "screens off" the association between C and S, if $\mathrm{Prob}(C/S\&G) = \mathrm{Prob}(C/G)$, and $\mathrm{Prob}(C/S\&{-G}) = \mathrm{Prob}(C/{-G})$. That is, the genetic predisposition would screen off the association between smoking and cancer if, among people who had the genetic predisposition, cancer were no more frequent in the smokers than in the nonsmokers and if the same were true among those who lacked the genetic predisposition.

If some G did so screen off cancer from smoking, then clearly the cigarette companies would be right, and the correlation between smoking and cancer would merely be due to the fact that smoking was more frequent among those who were going to get cancer anyway. Conversely, if there isn't any such G to explain away the original association then it seems to follow that there really is a direct causal link between the two. So let us take the statistical-relevance account to hold that A causes B just in case $\mathrm{Prob}(B/A) > \mathrm{Prob}(B)$ and there is no C which screens off B from A.

Why didn't the "standard view" need a "no screener-off" requirement? That is, why wasn't the standard view open to the objection that the probabilities appearing in its explanations might be "spurious" ones which made mere symptoms appear as causes? It is true that the standard view required high probabilities where the statistical-relevance view required only increases. But high probabilities can be spurious just as much as low ones.

The answer is that the standard view has already ruled out screeners-off by imposing the homogeneity requirement. The homogeneity requirement says that no further factors should make a difference to the probability of B. And so *a fortiori* it rules out any factors that will, so to speak, change $\mathrm{Prob}(B/A)$ from being bigger than $\mathrm{Prob}(B/{-A})$ to being equal to it.

But it is worth noting, and it will be important for what follows, that the converse implication does not hold. Although the homogeneity requirement implies the no-screeners-off requirement, the no-screeners-off requirement does not imply homogeneity. The reason is that the no-screeners-off requirement rules out only a quite specific kind of probability-altering factor, namely, those factors that, in altering the probability of *B*, make the association between *B* and *A* "disappear." So the nonexistence of screeners-off still leaves it open that there may be other probability-altering factors in the offing and that the homogeneity requirement is therefore not satisfied.

Thus, to return to our original example, although being old altered the probability of recovery and so stopped the simple penicillin/no-penicillin partition from being homogeneous for recovery, it doesn't follow that it screens recovery off from penicillin. It's still perfectly possible (and indeed no doubt true) that among the old, as among the young, recovery is more likely for those who take penicillin. And consequently the fact that being old alters the probability of recovery does not, on the statistical-relevance account, discredit the suggestion that the penicillin caused the recovery, when somebody who takes penicillin gets better.

Note that this now gives us something of a test case. I took it, in the first section, that when an old person who took penicillin recovered, we wouldn't be able to explain this. After all, what actually happened was terribly unlikely, given a full specification of the circumstances. But the statistical-relevance account disagrees. Provided that being old doesn't actually screen off the recovery from the penicillin, provided that the chance of recovery is still increased by taking penicillin, then, even if the probability of recovery remains very low, when we do get a recovery, we should deem it to be caused by the penicillin.

This makes it clear why the standard view requires homogeneity, where the statistical-relevance view is happy to settle for no screeners-off. The standard view says we have an explanation if all the factors present in the circumstances give the explanandum a high probability. Obviously all relevant factors need to be taken into account to determine whether this is so. But the statistical-relevance account doesn't put any demands on the absolute value of the explanadum's probability. All it requires is that this probability be increased by the presence of the putative cause, and that this increase not be screened off. And so the statistical-relevance account can safely ignore the question of whether any other factors have a similar influence.

However, though it is clear enough what is in dispute between the standard and statistical-relevance views, it is not at all clear which view is right. Indeed the situation seems downright puzzling. Even in the test case, *both* views seem highly plausible. Intuitions pull strongly in both directions. Surely unlikely events aren't causally explicable. But isn't an increase in probability *per se* enough to indicate causation? This tension should make us suspect that there is more to the dispute than meets the eye, that underneath the surface of the discussion a number of different issues are present. So indeed I shall shortly argue. But in order to do this I first need to elaborate a yet further view on the connection between probabilities and causes.

III. THE EVIDENTIAL VIEW

So far I have been implicitly presupposing that, whatever the exact form of the relationship between probabilities and causes, causal relationships are somehow *constituted* by probabilistic relationships. That is, I have intended that both the standard and the statistical-relevance views be understood to be claiming that one event's causing another just *is* a matter of there being certain probabilistic generalizations relating events of those types, analogously to the way that on more old-fashioned views causation is a matter of there being certain deterministic generalizations relating the event types. Henceforth I shall call views of this kind *definitional* views. I want to contrast definitional views with an *evidential* approach to the relationship between probabilities and causes.

(Perhaps it is implausible that anybody nowadays might suppose that causation can be *defined* in terms of probabilities, or indeed in terms of anything else. But, as will become clear shortly, it's not really the possibility of a full analysis of causation that's at issue, so much as opposing views on the connection between probabilities and the substance of causation.)

To illustrate the evidential view, let us pretend for a while (indeed for the rest of this section) that the "old-fashioned" deterministic view of causation referred to above is in fact correct: let us pretend that causation is necessarily deterministic. I want to show how, even given this assumption, probabilities (that is, probabilities other than 0 and 1) can tell us about causes. The idea will be that such probabilities can still be seen as *evidence* for the underlying presence of such deterministic relationships. (The assumption that causation is deterministic is not essential to the evidential approach. I return to this in the next section.)

It will be convenient to adopt John Stuart Mill's terminology for analyzing deterministic causation, and call some factor *A* a *cause* of some *B* just in case it is one of a set of conditions that are jointly and minimally sufficient for B. In such a case we can write $A\&X \rightarrow B$. In general there will also be other sets of conditions minimally sufficient for B. Suppose we write their disjunction as Y.

If now we suppose further that B is always determined when it occurs, that it never occurs unless one of these sufficients sets (let's call them B's *full causes*) occurs first[5], then we have

$$A\&X \vee Y \leftrightarrow B$$

Given this equivalence, it is not difficult to see why A's causing B should be related to A's being correlated with B. If A is indeed a cause of B, then there is a natural inference to $\text{Prob}(B/A) > \text{Prob}(B/-A)$: for, given A, one will have B if either X *or* Y occurs, whereas without A one will get B only with Y. And conversely it seems that if we do find that $\text{Prob}(B/A) > \text{Prob}(B/-A)$, then we can conclude that A is a cause of B: for if A didn't appear in the disjunction of full causes which are necessary and sufficient for B, then it wouldn't affect the chance of B occurring. This then gives the basic idea of the evidential view. But of course it is a bit too quick. To take the latter inference first, A's being a cause of B isn't the only way in which it can get to be correlated with it; for even if A is not itself a member of a full cause, it may be statistically associated with factors that are, and this in itself will give rise to a ("spurious") correlation between A and B. And conversely, if not so obviously, even if A *is* a cause of B it may be (though it would be something of a freak) negatively associated with elements of other full causes of B to just the extent required to leave us with a null correlation between A and B.

In effect what we are worried about here is the existence of factors that might screen off the correlation (or lack thereof) between A and B. But the evidential view can cope with this possibility by taking into account a more complicated set of statistical relationships. Suppose $[D]$ is the set of all the other causes of B which are associated with *A*. What we need to do is partition the original reference class by all combinations of presence and absence of the factors in [*D*], and then consider whether *A* is correlated with *B* *within* any of the cells of this partition. If *A* is so correlated then it must now be a cause, because there aren't any "remaining" associations within any cell left to produce "spurious" correlations between *A* and *B*. And, conversely, if there aren't any correlations between *A* and *B* left within cells, if any original correlation between *A* and *B* has been screened off by some member of [*D*], then it follows that *A* really isn't a cause after all.

---

[5] It is worth being clear about what is involved in assuming that some B is in this sense determined. We are assuming that there is, so to speak, determinism *in between* $A\&X \vee Y$ and B. This does not commit us to the view that the whole universe is deterministic. This point is important for the status of such probabilities as $\text{Prob}(B)$, $\text{Prob}(B/A)$, etc. I shall want to continue talking about such probabilities as general facts involving event types, and I shall want it to be possible for these probabilities to be other than 0 or 1 even when B is determined on each occasion of its occurrence. It is arguable, however, that serious probabilities other than 0 and 1 need to be underpinned by real indeterministic chances. But I don't need to take issue with this: we can still get serious nondegenerate probabilities even if B is determined by A, X, Y, etc., provided only that the occurrence of A, or X, or Y, etc., is itself the outcome of some real chance process, or that they are in turn determined by factors whose occurrence is the outcome of some chance process . . .

For those who would like to look more closely at these inferences there is an appendix that spells out the steps involved. Perhaps one brief observation will be helpful at this point. Partitioning by all *associated* causes [*D*] is not the same as partitioning by all other causes *tout court*. If the latter was required then, given the determinism we are assuming, we wouldn't be in a position to make inferences about causation until we had a partition that gave *B* a probability of either 1 or 0 within every cell. But since we don't have to control for all other causes—since, that is, taking all possible screeners-off into account doesn't require that we have a fully homogeneous partition—it *is* possible to make cogent causal inferences from probabilities other than 0 or 1.

IV. ADJUDICATION

a. *Deterministic Situations*. So now we have three different approaches: standard, statistical-relevance, and evidential. Which is right?

On the surface, perhaps, the statistical-relevance and the evidential accounts look similar. They both focus on increased probabilities rather than high ones. But in intent they are quite distinct. We can best bring this out by considering once more the kind of deterministic situation assumed in the last section. Thus let us suppose, again for the sake of the argument, that lung cancer (*C*) is always determined when it occurs, either by smoking (*S*) in conjunction with certain (presumably metabolic) preconditions *X* or by certain other conjunctions of factors *Y* which don't include smoking. It might seem that there will be a stand-off between the statistical-relevance and the evidential approaches in such cases. Both sides will agree about the structure of probabilistic and deterministic generalizations relating *C*, *S*, *X*, and *Y*. Both will agree that *S* causes *C*. But where the statistical-relevance approach will say that this is because *S* is correlated with *C* and this correlation isn't screened off by anything else, the evidential approach will say that this is because *S* is part of a full deterministic cause of *C* (and that the statistical relationships, including the absence of screeners-off, are merely indirect evidence for this).

It doesn't look as if there is much in it. But the statistical-relevance approach is in fact in trouble here. Consider an actual person who smokes and then gets cancer. According to the statistical-relevance approach, the smoking caused the cancer. But what if that person wasn't *X*, but was *Y*? It seems clear that in that case (let's call it the "case of the misleading cigarettes") the smoking wouldn't have caused the cancer—that though the person smoked, that wasn't why he or she got cancer.

And this of course is what the evidential approach will say. The probabilities are only indirect evidence for the underlying causal connection, and so smoking and cancer can be probabilistically connected in general without being causally connected in the particular case.

But the statistical-relevance approach seems to lack the room to say this. If causal connections are *constituted* by the relevant probabilistic associations, then it seems to follow from the general connection between smoking and cancer that the former causes the latter in the misleading particular case.

This case of the misleading cigarettes brings out an ambiguity that I have been slurring over so far. One can read '*A* causes *B*' as a relationship between types: 'being *A* is the kind of thing that causes *B*', 'on some occasions *A* gives rise to *B*', '*A can* cause *B*'. But one can also read it as a relationship between tokens: 'on this occasion *A*'s being instantiated caused *B*'s being instantiated'. In many contexts the distinction is unimportant—the token relationships between instantiations of *A* and *B* will obtain just in case the type relationships between the kinds as such obtain.

But for the evidential approach the distinction does matter: the type relationship that it takes to be evidenced by the probabilities (that *S* is a member of a set of conditions sufficient for *C*) doesn't guarantee that the relevant token relationship will obtain whenever a particular *S* is followed by a particular *C* (as when some other members of the sufficient set in question are absent). Which of course is why the evidential approach can deal with the misleading-cigarettes case.

But a corresponding move does not seem open to defenders on the statistical-relevance approach. By viewing the probabilities constitutionally rather than evidentially, it seems that they will be forced to conclude that a causal relationship between tokens obtains whenever we have tokens of kinds related by the appropriate probabilities.

Perhaps defenders of the statistical-relevance approach will feel inclined at this point to appeal to a distinction between type and token *probability* relationships. Of course, they will agree, in a general (type) sense smoking increases the probability of cancer. But it doesn't follow that it increases the probability in the particular case at issue, since, after all, the probability in the particular case was already 1, because of the presence of *Y*, quite independent of the smoking. Thus if we tie causation to token probabilities rather than type probabilities, we can avoid saying that the person with the misleading cigarettes got cancer because of smoking.

But note that this ploy means that the statistical-relevance approach has abandoned the distinctive no-screeners-off requirement. The switch from type probabilities to token probabilities, to probabilities that reflect the real chance of the outcome given all the circumstances to hand in the particular case, demands that our probabilities come from *homogeneous* reference classes, not just from classes that take into account all screeners-off.

And this then means that, in the deterministic case, the statistical-relevance approach collapses into the standard one, with both demanding that satisfactory causal explanations take all relevant factors into account. Indeed in the deterministic case both then leave us with the outcome always having a probability of either 0 or 1. And so neither the statistical-relevance nor the standard account has anything to say about deterministic situations which was not already said by the orthodox deductive covering-law model.[6]

b. *Indeterministic Situations*. Let me turn now to indeterministic situations. Some readers will no doubt feel this is long overdue. Perhaps the evidential view is the view demanded by deterministic assumptions. But the world isn't in fact deterministic. And so isn't the important quesiton whether the standard or the statistical-relevance view deals best with indeterminism?

But this would be too quick. The evidential view has things to say about indeterministic situations too.

This might seem a bit puzzling. In order to explain it we need to do a bit of unraveling. Leave the evidential view out of it for a moment. I have been speaking for the most part as if there were two definitional views of causation, the standard and statistical-relevance views. But we have also at various points come across what is in effect a third definitional view, namely, the "old-fashioned" deterministic view, according to which causation necessarily requires a deterministic generalization linking the event types involved. So from now on let us take it that there are three definitional views in the field: two indeterministic (the standard and statistical-relevance) and one deterministic ("the old-fashioned").

---

[6] Couldn't the statistical-relevance approach (or, for that matter, the standard approach) maintain that we could have explanations underpinned by nondegenerate probabilities in what were actually deterministic situations as long as we believed (with good reason?) that our probabilities were homogeneous? But though we would then "have explanations" insofar as we believed we could explain, there is a clear sense in which they wouldn't necessarily *be* explanations—for they would be downright mistaken whenever we had cases like that of the misleading cigarettes.

The evidential view is not in competition with definitional views. It is a view not about the constitution of causation, but about the possibility of a certain kind of indirect evidence for causal relationships. In section III, it is true, I introduced the evidential view by putting it in harness with the old-fashioned deterministic view of the constitution of causation. But, as I said at the time, this link was not essential. What I now want to show is that one might still need to adopt an evidential attitude even if one is taking an indeterministic view of the constitution of causation.

To see this, let us temporarily assume, for the sake of argument, such an indeterministic view of the constitution of causation. And let us agree that cancer, say, is not a deterministic outcome of prior circumstances. Even then, I want to argue, one needs to recognize the possibility of taking an evidential attitude to the connection between probabilistic generalizations and causal conclusions.

For it certainly seems intuitively clear that, even with these assumptions, one will want to allow that probabilities can tell us about causes even though the reference classes from which we are getting the probabilities are not homogeneous—surely we want to allow that we can have good grounds for thinking that smoking causes cancer even though the probabilities we take to show this would be altered by the introduction of further factors. Of course we need to recognize that here, as always, a sound inference requires that we take note of *some* of the further factors relevant to getting cancer. For we need to take into account possible screeners-off (which would include such factors as genetic make-up, psychological stress, diet, etc.). But this isn't the same as needing to take into account *all* relevant factors. No doubt there are as yet unknown metabolic differences between people which affect the chance of (without necessarily determining) cancer. But provided that we have reason to believe that these differences aren't themselves correlated with smoking, as they needn't be, and so can't screen off cancer from smoking, our ignorance of what they are and what exact effect they have on the probabilities clearly doesn't invalidate the inference to a causal conclusion.

The importance of this point, that even in the indeterministic situation inferences to causes demand no screening off rather than homogeneity, is that it shows that the evidential view is still at work. Consider once more the case of the misleading cigarettes. As originally described, this presupposed determinism. But we can get an indeterministic analogue. Let smoking plus *X*, and the various factors in *Y*, give a homogeneous partition for cancer, but now one that doesn't always determine cancer or its absence, but merely makes it probable to various degrees. Then one might have a particular person who smoked, and had a raised chance of cancer, and got cancer—but not because of the smoking, for they lacked *X*, but because of *Y*. Although smoking raises the probability of cancer, and this association isn't screened off, it doesn't follow that when a smoker gets cancer the smoking caused the cancer.

This shows that if you are concerned with the realization of causation in particular cases, with information that will tell how the causes are constituted, then, as in the deterministic situation, you are going to need homogeneous reference classes. It is only when you have such homogeneous probabilities that you can be sure that the factors you allude to by way of explanation in the particular case are those which in *that* case made the result more probable, not just factors that have the appropriate general (type) relationship with the outcome.

And, conversely, this shows that if you are happy with merely unscreened-off probabilities, rather than homogeneous ones, then you must be adopting an evidential rather than a definitional view of the relationship between probabilities and causes. The probabilities can't then be the substance of the causal relationships, but merely an indication of underlying connections.

It might still be felt that I am avoiding the main issue. Perhaps definitional views of causation require homogeneous probabilities across the board, in indeterministic as well as deterministic situations, and perhaps to that extent the statistical-relevance view should never have mentioned the weaker no-screeners-off requirement in the first place. Perhaps, that is, the no-screeners-off idea is of importance only to the evidential view. But doesn't there still remain the original question: what is the right definitional account of causation for indeterministic situations? That is, even if indeterministic accounts—the standard and statistical-relevance accounts alike—require homogeneous probabilities, isn't there still the question of whether these probabilities should be high probabilities or merely increased ones?

There is, however, a good reason why I have spent so long on the evidential view before turning to this question, namely, that once we have distinguished the evidential view, and seen that it can apply in both deterministic and indeterministic situations, there is far less reason than before to suppose that *either* the statistical-relevance view *or* the standard view is correct.

Let me deal first with the statistical-relevance view. What I want to suggest is that the statistical-relevance view seems plausible in the first place only because it gets confused with the evidential view.

The statistical-relevance view implies that somebody who has only a 10% chance of getting cancer, given *all* the relevant factors present in the particular situation, can be caused to get cancer by those factors. I find this counterintuitive. Just this person might well (nine times in ten) not have got cancer. Surely then we can't explain why he or she did?

This was the intuition I appealed to in section I when I first argued that the maximal-specificity requirement needed to be added to the standard view. But this intuition is rather unfashionable nowadays. And of course the reason is the further thought, familiar from the writings of statistical-relevance theorists, that, if smoking increases the probability of cancer, even if only from 1% to 10%, then this in itself indicates that smoking causes cancer.

But I think we should resist fashion and stand by the original intuition. For I think fashion has been misled by a failure to distinguish the statistical-relevance from the evidential view, and has therefore misread the significance of increased probabilities. It is right to hold, with the evidential view, that increased probabilities are important. But it is wrong to conclude, with the statistical-relevance view, that they can constitute causation.

According to the evidential view, increased probabilities are exactly what we need in order to *infer* the existence of underlying causal connections. That smoking increases the probability of cancer from 1% to 10% indeed gives us good reason to conclude that smoking causes cancer. But on the evidential view we don't need to suppose that the causation itself is *constituted* by a mere 10% chance. The conclusion of our inference, on this reading, could be that there existed certain other unknown factors which, when they occurred together with smoking, gave cancer a high probability.

Thus the evidential view allows us both to stand by the original intuition that causation requires high probabilities and to accommodate the insight that increased probabilities are important. And it seems to me that, once this is recognized, the statistical-relevance view loses its plausibility. Once we can give a nondefinitional account of the feeling that increased probabilities are important, there remains little to be said in favor of the statistical-relevance view.

What now of the standard view? Here things are less clear. It might seem that if the statistical-relevance view is out of the running, then the standard view must be the right account of indeterministic causation. However, we shouldn't take it for granted in the present context of argument that there is any "right account" of indeterministic causation. Perhaps, that is, all indeterministic accounts of causation should be rejected, in favor of the old-fashioned deterministic account.

It is important to recognize here that the question of whether *causation* is definitionally indeterministic or not is in principle quite independent of whether the *world* is indeterministic or not. There is every reason to believe that the world is indeterministic. But one can allow this, and still stand by a deterministic view of causation. All that would follow would be that, to the extent that things aren't determined, they aren't caused either.[7]

What arguments are there for allowing indeterministic causation? I suspect that many people have been swayed by the following line of thought. "It is scarcely deniable that probabilities have *something* to do with causes. A huge amount of empirical research, by medical investigators and agricultural scientists, not to mention psychologists, economists, and sociologists, proceeds from probabilistic data to causal conclusions. Clearly this isn't all just based on a conceptual mistake. So surely we should accept that causes can be constituted by probabilities as well as by deterministic certainties." But the evidential view shows that this reason for rejecting the old-fashioned deterministic view of causation is mistaken. For the evidential view, as originally illustrated in section III above, shows how it would still make perfect sense to infer from probabilities to causes even if one assumed that causes were always constituted by underlying deterministic connections.

But this scarcely decides the matter. Perhaps there are independent reasons, apart from the cogency of basing causal conclusions on probabilistic data, for thinking causation needn't be deterministic. And on reflection it does perhaps seem implausible to say that indeterminism *per se* precludes causation. Suppose that roulette wheels are really indeterministic (suppose, that is, that quantum effects always have an influence on where the ball lands). Should that stop us saying (on most evenings) that the house is caused to win by the presence of a zero on the wheel? (Which, note, gives it an extremely high probability of winning, not just a higher probability than it would otherwise have.)

---

[7] One could even, though there wouldn't be much point, do it the other way round: one could combine an acceptance of an indeterministic definitional view of causation with the thesis that the world is in fact deterministic, holding then that even if the world had been indeterministic, such indeterminism wouldn't have implied an absence of causation.

But neither is this reflection conclusive on the other side. The old-fashioned view doesn't need to deny that there is causation in the casino. If roulette wheels are really indeterministic, then the physical nature of the set-up determines that there is a certain single-case probability (a *chance*) of the house winning. And so the old-fashioned view can say that what is caused in the casino is the very high chance of the house winning over the evening. This causation is itself deterministic. And that, so the old-fashioned thought would go, is all the causation there is. Once the high chance has been explained, there is no further explanation of why the house won—after all, given just the same chance on another night, the house might not have won, so why suppose we can explain its winning, as opposed to its being very likely that it would?

It might seem that there is little to choose at this stage between the standard and old-fashioned ways of admitting causation into the casino. However there is a strong reason, given what has gone before, for preferring the old-fashioned account. This relates once more to the possibility of drawing causal conclusions from inhomogeneous but screener-off-free probabilistic data.

Note that once we relax deterministic assumptions, we cannot continue to assume that an unscreened-off $\mathrm{Prob}(B/A) > \mathrm{Prob}(B)$ shows that A is (part of) a determining condition for B. What it shows (as the appendix explains) is rather that A is part of a homogeneous reference partition for B. But—and this is the crucial point here—this doesn't mean that A is (part of) a set of factors that makes the chance of B *high*. And so, if the standard view were right, and causation required that A (plus some X) made the chance of B high, then there would be no valid inference from an unscreened-off $\mathrm{Prob}(B/A) > \mathrm{Prob}(B)$ to the conclusion that A is a cause of B.

There are two ways to go here. (1) We could reject the assumption that unscreened-off correlations indicate causal conclusions; or (2) we could insist on the old-fashioned view in place of the standard view, and maintain that in indeterministic situations it is only the chances of undetermined events that get caused, rather than the events themselves. I take it that (1) is unacceptable—intuition, instantiated in a whole tradition of empirical research, tells us that *something* about causal conclusions must follow from un-screened-off correlations. So I maintain (2), saying that when we have an un-screened-off $\mathrm{Prob}(B/A) > \mathrm{Prob}(B)$ in an indeterministic situation, what we can infer is that A is (part of) a deterministic cause of the chance of B.[8]

This rejection of the standard view for the old-fashioned view might seem a somewhat unsatisfactory maneuver forced on me only because of my rejection of the statistical-relevance view. For note that the statistical-relevance view, unlike the standard view, does allow us to draw causal conclusions from un-screened-off correlations in indeterministic situations: increased probabilities, by contrast with high ones, are indeed reliably indicated by un-screened-off correlations.

Even so, I think it is right to reject the statistical-relevance view. As I have argued, the idea that mere increase in probability constitutes causation is in itself counterintuitive; and what plausibility there is to the statistical-relevance view can be explained away as due to its conflation with the evidential view. We now see that the alternative to the statistical-relevance view is to construe causation in indeterministic situations as the deterministic fixing of chances by prior circumstances. Even if this seems awkward at first sight, I don't see that there is anything definite to be said against it. And so I see no reason not to stand by the rejection of the statistical-relevance view, and to return to the old-fashioned deterministic view of causation across the board.

DAVID PAPINEAU
University of Cambridge

APPENDIX

Let me start by going back to the assumptions of Section III. That is, let me take it not only that causes are deterministic, but also that B is determined whenever it occurs (so we can write: $A \& X \vee Y \leftrightarrow B$). I want to show that, given these assumptions, A will be a cause of B if and only if there is an un-screened-off probabilistic association between A and B.

This argument has two stages. I shall first consider the simple case where A is probabilistically independent of X and Y and truth functions thereof. Then I shall proceed to the more complicated case where such independence is not assumed.

---

[8] I would like to thank Crispin Wright for help with the argument of the last two paragraphs.

The simple case. Note that

$$\begin{aligned}
\mathrm{Prob}(B/A) &= \mathrm{Prob}(A\&X \vee Y/A)\\
&= \mathrm{Prob}(A\&X \vee A\&Y)/\mathrm{Prob}(A)\\
&= [\mathrm{Prob}(A\&X) + \mathrm{Prob}(A\&Y) - \mathrm{Prob}(A\&X\&Y)]/\mathrm{Prob}(A)\\
&= \mathrm{Prob}(X) + \mathrm{Prob}(Y) - \mathrm{Prob}(X\&Y)\\
&\text{(since } A \text{ is independent of } X,\ Y, \text{ and } X\&Y)\\
&= \mathrm{Prob}(X \vee Y)
\end{aligned}$$

So if $\mathrm{Prob}(B/A) > \mathrm{Prob}(B)$, then $\mathrm{Prob}(X \vee Y) > \mathrm{Prob}(B)$. That is, $\mathrm{Prob}(X \vee Y) > \mathrm{Prob}(A\&X \vee Y)$. Which means that there is indeed some X that in conjunction with A figures in the list of B's full causes. [If there weren't—if, so to speak, that X were the impossible condition—then $\mathrm{Prob}(X \vee Y)$ would have to equal $\mathrm{Prob}(A\&X \vee Y)$.] Thus $\mathrm{Prob}(B/A) > \mathrm{Prob}(B)$ implies that A is a cause of B.

Conversely, if A is a cause of B—if there is a "non-impossible" X such that $A\&X$ figures in the list of B's full causes—then $\mathrm{Prob}(B/A)$, which is equal to $\mathrm{Prob}(X \vee Y)$, must be bigger than $\mathrm{Prob}(B)$, which equals $\mathrm{Prob}(A \& X \vee Y)$. The only way we could have an equality would be if $X \vee Y$ never occurred in the absence of A, but this is ruled out by the independence assumption. [A slight complication. If A *always* occurred, then $X \vee Y$ would never occur in its absence, even if $X \vee Y$ was probabilistically independent of A, and we would have $\mathrm{Prob}(B/A) = \mathrm{Prob}(B)$. Perhaps we should say that something that "always" occurs, whatever precise modal significance we attach to this, cannot qualify as a cause. But what then about something that "almost" always occurs, in the measure-theoretic sense? Again, this will give $\mathrm{Prob}(B/A) = \mathrm{Prob}(B)$. Here I think we should simply relax the evidential thesis, and admit that to this extent causes won't automatically show up in probabilities.]

I turn now to the *second stage* of the argument, where we relax the independence assumption that A is independent of X and Y and truth functions thereof. Now we need to "control" for the elements in X or Y (remember that X and Y will themselves generally be disjunctions or conjunctions of factors) which are associated with A and which therefore might be "confounding" the probabilistic relationship between A and B. That is, we need to look separately at classes of cases similar with respect to the presence or absence of such confounding elements, and see whether A is still probabilistically associated with B within such classes. The claim, then, is that A is a cause of B if and only if such "un-screenable-off" associations remain.

The strategy here is simple enough. The original reference class is partitioned by means of combinations of presence and absence of factors from X and Y, in such a way that A is independent of the "remaining" parts of X and Y and truth functions thereof within each cell of the partition. Then in effect we repeat the original argument within each cell of the partition. Given the assurance that within each cell A is independent of the "remaining" causes, we can conclude that (a) if there are some cells in which *A* is still probabilistically associated with *B*, then it must be a cause (there must be some condition together with which *A* appears on the list of *B*'s full causes), and (b) if there are no cells in which *A* remains probabilistically associated with *B*, then it can't be a cause (modulo the qualification made earlier about *A*s which "almost" always occur, now relativized to cells). If *A* is associated with *B* within some cells but not within others, then *A* and the factors defining those cells are said to "interact" in causing *B*. This will occur if *A* is associated with (and we therefore have to "control" for) some factor which is a necessary part of *X*.

All the above argument is set within the deterministic assumptions of Section III. To generalize the explication of the evidential approach to indeterministic situations, we simply need to "replace" all undetermined events with their chances of occurrence, and then proceed as before. That is, we can regard the aim of the evidential approach in indeterministic situations to be the identification of the *deterministic* laws relating the chances of undetermined events to antecedent conditions. The presence of an un-screened-off correlation between *A* and *B* now indicates that *A* is one of the conditions that plays a part in determining the chance of *B*. Note that this does not mean that *A* plays a part in determining that *B* is *highly* probable, for the relevant chance of *B* need not be high.

DP

TRANSLUCENT BELIEF*

ONE of the more perplexing problems in the philosophy of language involves the interpretation of belief ascriptions—sentences of the form:

R believes that *p*.

where the expression that replaces *p* (the content of the ascription) has the syntactical form of a sentence. The problem is to say how such contents are to be interpreted.

Philosophical lore has it that our alternatives are exactly two: either an ascription of belief is transparent, or it is opaque.[1] If it is transparent, the content is just the sentence it appears to be, and its interpretation is straightforward. If it is opaque, this is not the

---

*I am grateful to Israel Scheffler, Nelson Goodman, and Jonathan Adler for helpful comments on an earlier draft of this paper, and to the University Research Council of the University of North Carolina at Chapel Hill for financial support.

[1] Some may be one, some the other. Indeed, some tokens of a single type may be transparent, others opaque. But tradition has it that unless a token is transparent, it is opaque. There is no third alternative.

0022-362X/85/8202/0074$01.70

case. Philosophers disagree about how opaque constructions are to be interpreted, but this much is generally agreed: expressions occurring in opaque contexts do not have the same referents as they have in transparent contexts; and the wording of an opaque construction affects the truth value of the sentence that contains it in ways in which the wording of a transparent construction does not. In a transparent context we, as it were, look through our words to their referents; in an opaque context we attend to the words themselves.

The conviction that the disjunction "transparent or opaque" exhausts our options strikes me as mistaken. For many ascriptions of belief, neither construal is satisfactory. An intermediate position is wanted—one in which due attention is paid both to the words that are the medium of reference and to the things that are the objects of reference. I suggest then that we extend the metaphor and say that ascriptions of belief are typically neither transparent nor opaque, but *translucent*. A medium is literally translucent if it transmits light, but diffuses it in such a way that the outlines of objects cannot be clearly discerned. Correspondingly, a linguistic construction is metaphorically translucent if it transmits reference, but in such a way that the limits on paraphrase are not obvious. Just as we must attend to the nature of a literally translucent medium to understand what can be seen through it and what distortions and obscurities are to be expected, we must do likewise in the metaphorical case to understand how reference is transmitted, distorted, and obscured in ascriptions of belief. Just as literally translucent media range from the all but transparent to the all but opaque, so do metaphorically tranluscent media. And just as a translucent medium may block some light waves and transmit others, so a metaphorically translucent medium may block some substitutions but permit others. We should expect the limits on permissible paraphrase in translucent contexts to vary considerably—some such contexts being more nearly transparent, others more nearly opaque.

My strategy in this paper is as follows: I review familiar accounts of the interpretation of belief ascriptions to identify their characteristic strengths and weaknesses. I then formulate an account in which such ascriptions are construed as translucent. I suggest that it retains important virtues of its predecessors and avoids their major failings. With suitable modifications, my account can be extended to cover ascriptions of other so-called "propositional attitudes"—fears, hopes, expectations, and the rest. But I will not attempt that project here. I should emphasize that it is no part of my project to say what beliefs are. I am concerned exclusively with the interpretation of expressions that follow the phrase 'believes that' in ascriptions of belief. I think significant progress can be made on that problem without settling the reference of the term 'belief'—without, that is, solving substantive problems in the philosophy of mind.

When belief ascriptions are construed transparently, the terms in their content statements have their normal referents.[2] This allows for an easy understanding of some of the ways beliefs are said to function in epistemology. We classify beliefs as true or false, as confirmed, disconfirmed, or unconfirmed, as rational or irrational, and so on. The belief ascribed to Sam in

(A) Sam believes that kangaroos are carnivores.

is true only if kangaroos eat meat; it is confirmed only if there is evidence that they do so; it is rational only if there is sufficiently good reason to think that they do so; etc. Information about the dietary habits of kangaroos is plainly germane to the epistemic status of Sam's belief if in the statement of that belief we make reference to kangaroos. But if in the context of belief ascriptions 'kangaroo' does not refer to kangaroos, it is hard to see why or how such animals should matter to the epistemic evaluation of Sam's belief. It is then a virtue of the transparent construal that it preserves a clear connection between the interpretation of belief ascriptions and the epistemic status of the beliefs they ascribe.

In transparent contexts, all coextensive terms are intersubstitutable *salva veritate*. So it follows from (A) that

(B) Sam believes that marsupials of the family macropodidae are carnivores.

even if Sam, being ignorant of taxonomy, is in no position to affirm (B). Indeed, for every belief a person ascribes to himself, indefinitely many other ascriptions are equally applicable—namely, all those which result from the substitution of coextensive terms for the terms in his ascription. Many of these he will not recognize as statements of his belief; indeed, many he will not even understand. Moreover, all such paraphrases are on a par. Those whose truth he acknowledges have no special standing.

Although Sam is prepared to say that kangaroos are carnivores, he is under the misapprehension that carnivores are eaten as meat, not eaters of meat. As regards the eating habits of kangaroos,

(C) Sam believes that kangaroos are not meat eaters.

---

[2] Or at least their occurring in the context of a belief ascription does not affect their reference. Metaphor, ambiguity, vagueness, deviant usage may occur in belief ascriptions just as in any other sentences.

His willingness to accept the contents of both (A) and (C) shows Sam to have incompatible beliefs about the place of kangaroos in the food chain. Since he is unlikely to be in a position to feed or to be fed to kangaroos, this seems but a minor defect in his doxastic system.

One of the strengths of a transparent construal is that it enables us to ascribe mutually incompatible beliefs to an agent. Although we often do so, this is apt to be overlooked in the formulation of theories of belief ascription. The conviction that it is always a mistake to harbor incompatible beliefs slides easily over into the conviction that it is always a mistake to ascribe them.[3] But since we regularly ascribe such beliefs, our system should have the resources to say so. And since people actually harbor such beliefs, it should not follow from our system that such ascriptions are inevitably in error. In these respects, the transparent construal of belief ascriptions appears satisfactory.

But a difficulty remains. On a transparent reading, (A) is equivalent to

(D) Sam believes that kangaroos are meat eaters.

So Sam's beliefs are not merely incompatible, they explicitly contradict each other. What looked to be a minor flaw in his factual beliefs now looks to be a serious logical failing.

Still, it seems wrong to convict Sam of an error in logic. He does not, after all, recognize the contents of both (C) and (D) as statements he believes. He is well aware that one of them must be wrong, even if he has no idea which. Moreover, instruction in logic will not avail him in recognizing and correcting his error. But instruction regarding the eating habits of kangaroos and carnivores, or regarding the interpretation of the terms 'kangaroo' and 'carnivore' might do so.[4] If, however, our ascriptions are transparent, we have no basis for calling his error zoological or semantical rather than logical. Epistemology requires a more sensitive device for delineating beliefs than the transparent reading provides.

Do opaque construals fare any better? In considering the matter we should distinguish between Quinean and Fregean accounts. The former take the content of a belief ascription to be semantically simple; the latter take it to be semantically complex.

Thus W. V. Quine[5] construes the phrase 'believes that kangaroos are carnivores' as an unbreakable one-place predicate that applies to Sam if (A) is true. 'Kangaroo' then no more occurs as a term in (A) than 'cat' occurs as a term in 'cattle'. This enables us to say that sentences (A)-(D) ascribe distinct, logically independent beliefs to Sam. But Quine's criterion is too sensitive in that any syntactic difference in the content of belief ascriptions marks a difference in the beliefs ascribed. And it is too insensitive in that syntactically identical sequences with different interpretations count as the same belief.[6] Attempts to evade the latter difficulty by relativizing ascriptions to a language or to someone's use of a language reintroduce the problematic semantic features that the account was designed to avoid.[7] Moreover, if belief contents are construed as uninterpreted syntactical sequences, it is not clear that they have any bearing on epistemology. Questions about the truth of, or evidence for, beliefs so construed receive no satisfactory answer.

---

[3] Quine's "principle of charity" encourages this tendency. See *Word and Object* (Cambridge, Mass.: MIT Press, 1960), p. 59.

[4] Not that it is always wrong to ascribe explicitly contradictory beliefs to someone. But such an ascription does not seem to capture the nature of Sam's error.

[5] *Word and Object*, pp. 211–216.

Perhaps what is wanted is an intermediate position—one in which such expressions are credited with interpretations distinct from but systematically related to the interpretations that they have in transparent contexts.[8] This is the sort of account that Frege seeks to provide. He contends that the referent of a term in an opaque context is the sense that term has in a transparent context.[9] (A) then contains no terms that refer to kangaroos or carnivores. But it contains terms that refer to the senses of 'kangaroos' and 'carnivores'. And since the sense of a term determines its reference, (A) is still related, albeit indirectly, to such animals.

In an opaque context, only terms with the same sense are intersubstitutable. If 'kangaroo' and 'marsupial of the family macropodidae' differ in sense, then (B) cannot be obtained from (A) by substitution. In that case (A) and (B) ascribe distinct beliefs to Sam.

Problems remain. Since it is not clear which expressions (if any) have the same sense, it is not clear which expressions are intersubstitutable in opaque contexts. Moreover, even if two expressions are in fact alike in sense, someone might mistakenly think otherwise.

---

[6] See Alonzo Church, "On Carnap's Analysis of Statements of Assertion and Belief," *Analysis*, x, 5 (April 1950): 97–99.

[7] Quine, of course, recognizes this and concludes that the vocabulary of propositional attitudes will have no role in limning the true and ultimate structure of reality. *Word and Object*, p. 221.

[8] See David Kaplan, "Quantifying In," in Donald Davidson and Jaakko Hintikka, eds., *Words and Objections* (Boston: Reidel, 1969), pp. 206–242. Although I agree with Kaplan about the need for an intermediate position, we disagree about how that need is to be satisfied.

[9] Gottlob Frege, "On Sense and Reference", *Translations from the Writings of Gottlob Frege*, Peter Geach and Max Black, eds., (Oxford: Blackwell, 1970), p. 59.

For, like the rest of our beliefs, our semantic beliefs are subject to error. If 'carnivore' and 'meat eater' have the same sense, (A) is equivalent to (D), whether Sam thinks so or not. Sam still stands convicted of harboring logically incompatible beliefs—those ascribed in (C) and (D)—even though he does not take himself to believe that kangaroos are meat eaters, and hence has no reason to think that he has any belief that conflicts with (C).

We might avoid this outcome by relativizing our interpretations to idiolects.[10] Since 'carnivore' and 'meat eater' have different senses in Sam's idiolect, (D) cannot be obtained from (A) by substitution. But upon learning that the terms are commonly taken to have the same sense, Sam is apt not only to revise his idiolect, but also to regard his previous understanding of the terms as mistaken. If, however, senses are determined by idiolects, he has no reason to do so. For in that case his usage is not mistaken, just idiosyncratic. And, given the senses the terms have in his idiolect, there is no conflict between (A) and (C). Sam's beliefs are not incompatible; his usage is simply unusual. Relativizing interpretations of opaque constructions to idiolects thus obliterates errors rather than explaining them. We seek an account that explains what is wrong with believing both that kangaroos are carnivores and that they are not meat eaters—not one that denies that there is anything wrong with doing so.

Frege's account is open to another objection. We regularly permit substitutions of terms whose senses plainly differ. For example, Max's state of mind with respect to a single arrest might be described as:

(E) Max believes that the brains behind the Brinks job has been arrested.

or as:

(F) Max believes that the Brooklyn bank robber has been arrested.

'The brains behind the Brinks job' and 'the Brooklyn bank robber' manifestly differ in sense. What allows for their intersubstitutability in statements of Max's belief is his conviction that they refer to the same individual. But it will not do to construe the context as transparent. For even though the fellow—S. J. Lawless, to give him a name—is also responsible for the Sixth Avenue sapphire snatch, Max does not know it. He considers that crime unsolved, and thinks its perpetrator is still at large. So although some identifying descriptions of Lawless can be freely substituted for one another in the statement of Max's belief, others cannot. (E) and (F) are statements of Max's belief in a way in which

---

[10] Such a move is decidedly un-Fregean. For Frege insists that senses are public, shared entities. "On Sense and Reference," p. 59.

(G) Max believes that the Sixth Avenue sapphire snatcher has been arrested.

is not. There are then nontransparent contexts in which sameness of sense is not the criterion for intersubstitutability of terms. Frege's account cannot accommodate these.

Many other theories of the interpretation of belief ascriptions have been proposed. But they are largely variations on themes I have already identified. Since my purpose here is to motivate consideration of an alternative approach, it is perhaps unnecessary to detail the ways in which I take the various neo-Fregean interpretations to come to grief.

I suggest that the interpretation of a belief ascription typically turns on both the classification and the truth conditions of the content sentence. In its dependence on classification, my account is reminiscent of opaque construals; in its dependence on truth conditions, it is reminiscent of transparent ones. It is my contention that by combining aspects of each we obtain a more adequate account than either alone provides.

Following Nelson Goodman and Israel Scheffler,[11] let us introduce the schema *p-sentence* whose replacements denote all and only those sentences which are paraphrases of the sentence that replaces *p*. Very roughly, a *p*-sentence is a sentence to the effect that *p*. A replacement for the schema *p-sentence* then is a general term that denotes sentences. 'Kangaroos-are-carnivores-sentence' denotes all and only paraphrases of 'Kangaroos are carnivores'. And 'kangaroos-are-carnivores-sentence' is coextensive with 'kangaroos-are-meat-eaters-sentence' if and only if the two terms denote exactly the same sentences.

The introduction of *p*-sentences poses no particular problem.[12]

---

[11] Goodman, "On Likeness of Meaning," *Problems and Projects* (Indianapolis: Hackett, 1972), pp. 221-230; Scheffler, "An Inscriptional Approach to Indirect Quotation," *Analysis*, XIV, 4 (March 1954): 83-90. Notice that what Scheffler, Goodman, and I do with *p*-labels is very close to what Wilfrid Sellars does with dot quotes. See his *Science and Metaphysics* (London: Routledge & Kegan Paul, 1968), pp. 80-90.

[12] It is sometimes argued that the introduction of terms by means of such schemata would require the language to have infintely many semantically independent primitives. Such a language, it is claimed, would be unlearnable. See Davidson, "Theories

We regularly recognize sentences as paraphrases of one another. And our judgments of pharaphrase can serve as a basis for classifying diverse sentences as instances of a common *p*-sentence. Ordinary standards of paraphrase, of course, do not yield necessary and sufficient conditions for the instantiation of a particular *p*-sentence. But this is a common feature of our understanding of general terms. We are equally without necessary and sufficient conditions for the instantiation of the term 'chair'. But we have no difficulty using the term or recognizing chairs when we encounter them. There seems no compelling reason to impose more rigorous standards on the terms we use to classify linguistic items than we impose on those we use to classify nonlinguistic ones.

We can begin our explication of (A) then by saying

(A) Sam believes that kangaroos are carnivores.

only if

(A′) Sam cognitively favors some kangaroos-are-carnivores-sentence.

I use the vague and unfamiliar term 'cognitively favors' to avoid unwanted associations suggested by more familiar mentalistic terminology. Without defining 'cognitively favors', I can say something about its application. Among the sentences an individual cognitively favors are those he is inclined to utter, inscribe, or assent to sincerely. These, presumably, are determined by his dispositions to verbal behavior, and the evidence that he favors them derives from the sentences he actually employs. Typically, however, his dispositions to verbal behavior do not exhaust the class of sentences he cognitively favors. If he is self-deceptive, insincere, lacking in self-awareness or in linguistic competence, or simply reticent, he will cognitively favor sentences he has no inclination to utter, inscribe, or assent to. So an individual may cognitively favor sentences without knowing it. The evidence that he does so comes from his nonverbal behavior, together with the equivocations, hesitations, hints, and hedges in the sentences—if any—that he employs. Both

---

of Meaning and Learnable Languages," in *Logic, Methodology, and Philosophy of Science* (Amsterdam: North-Holland, 1966), pp. 383/4. Actually, there is no problem. Even though the replacements of the schemata are one-place predicates, there is no reason to think they have to be learned one by one. See Robert Schwartz, "Infinite Sets, Unbounded Competences, and Models of Mind," in C. W. Savage, ed., *Perception and Cognition: Minnesota Studies in the Philosophy of Science* IX (Minneapolis: U of Minnesota Press, 1978), pp. 183–200; and Goodman, "Splits and Compounds," in his *Of Mind and Other Matters* (Cambridge, Mass.: Harvard, 1984), pp. 77–80.

verbal and nonverbal behavior then yield evidence of cognitive favoring—evidence that is used in our ascriptions of belief.[13]

Sentences are symbols in a language. If the syntactical sequence that an agent utters, inscribes, or accepts has different interpretations in different languages and if contextual cues are too meager to decide among the rival interpretations, we cannot tell which sentence he cognitively favors. Before we can use such a sequence as the basis for a belief ascription, we must settle on an interpretation.

Ascribing a belief to someone then involves identifying and classifying a sentence we take him to cognitively favor. That sentence need not be the one that replaces *p* in our schema. If Sam speaks only French, our basis for (A) might be his cognitively favoring

Les kangourous sont carnivores.

Nor need we know exactly which sentence he cognitively favors. All we need know is that he cognitively favors *some* sentence that is correctly paraphrased as 'Kangaroos are carnivores'.

A person's cognitively favoring a sentence normally reflects his commitment as to how things stand as regards the entities denoted by its terms. So ascribing a belief typically involves the imputation of such a commitment. This requires the addition of a truth condition. The belief ascribed to Sam in (A) is true just in case there are such things as kangaroos and whatever is a kangaroo is a carnivore. That is, just in case:

$$(\exists x)(x \text{ is a kangaroo} \mathbin{\&} (y)(y \text{ is a kangaroo} \supset y \text{ is a carnivore}))$$

If we incorporate this into our explication of (A), we get

$$(\text{A}'') \quad (\exists s)(\text{Sam cognitively favors } s \mathbin{\&} s \text{ is a kangaroos-are-carnivores-sentence} \mathbin{\&} (s \text{ is true} \supset ((\exists x)(x \text{ is a kangaroo} \mathbin{\&} (y)(y \text{ is a kangaroo} \supset y \text{ is a carnivore}))))$$

We cannot represent

Kangaroos are carnivores.

---

[13] These features enable the notion of cognitively favoring to function in an explication that comprehends not just beliefs that the agent can readily articulate, but also beliefs he does not acknowledge, and even ones that he does not recognize as his own. Moreoever, it allows for the ascription of beliefs to human infants and to nonhuman animals. An infant or an animal who displays sufficiently sophisticated differential behavior as regards a particlar situation can be said to cognitively favor a sentence to the effect that that situation obtains. See Quine, "Quantifiers and Propositional Attitudes," in *The Ways of Paradox and Other Essays* (New York: Random House, 1966), p. 194; and Ruth Barcan Marcus, "Rationality and Believing the Impossible," this JOURNAL, LXXX, 6 (June 1983): 332–333.

simply as

$$(x)(x \text{ is a kangaroo} \supset x \text{ is a carnivore})$$

for the latter is true if kangaroos don't exist. And Sam cannot plausibly be said to believe that kangaroos are carnivores without its committing him to the existence of kangaroos and carnivores.

This imputation of ontological commitment might seem to raise difficulties for the interpretation of fictional beliefs, especially if we admit no fictive entities into our ontology. But if we accept Goodman's account of fictive language,[14] there is no real problem. According to Goodman, what matters in the interpretation of such language is not what fictive terms denote, but rather what terms denote them. So the fictional sentence

Leprechauns are wee people.

is explicated as

Leprechaun-descriptions are wee-people-descriptions.

Accordingly,

(H) Pat (fictively) believes that leprechauns are wee people.

is to be explicated as

(H′) $(\exists s)(\text{Pat cognitively favors } s \mathbin{\&} s \text{ is a leprechaun-descriptions-are-wee-people-descriptions-sentence} \mathbin{\&} (s \text{ is true} \supset (\exists x)(x \text{ is a leprechaun-description} \mathbin{\&} (y)(y \text{ is a leprechaun-description} \supset y \text{ is a wee-people-description}))))$

Notice how this differs from the belief of the more gullible Mike:

(I) Mike believes that leprechauns are wee people.

The content of his belief is a false factual sentence. Accordingly, (I) is to be explicated as

(I′) $(\exists s)(\text{Mike cognitively favors } s \mathbin{\&} s \text{ is a leprechauns-are-wee-people-sentence} \mathbin{\&} (s \text{ is true} \supset (\exists x)(x \text{ is a leprechaun} \mathbin{\&} (y)(y \text{ is a leprechaun} \supset y \text{ is a wee person}))))$

The difference in the ontological commitments imputed to Pat and Mike are found in the truth conditions set out in (H′) and (I′), respectively. Pat's belief commits her to the existence of leprechaun-descriptions and wee-people-descriptions, whereas Mike's commits him to the existence of leprechauns and wee people. The descriptions exist, so the truth condition on Pat's belief is satisfied; but leprechauns and wee people do not, so the truth condition on Mike's belief is not.

---

[14] *Languages of Art* (Indianapolis: Hackett, 1976), pp. 21–27. See also my *With Reference to Reference* (Indianapolis: Hackett, 1983), pp. 43–50.

The truth condition, like the rest of the explication, is strictly extensional. If the predicates have their normal interpretations, the beliefs ascribed to Sam in (A), (B), and (D) have the same truth condition; for 'kangaroo' is coextensive with 'marsupial of the family macropodidae' and 'carnivore' with 'meat eater'. Sam, of course, need not know this. For nothing in the explication requires him to be cognizant of the truth conditions of the sentences he cognitively favors.

Indiscriminate use of the truth predicate produces paradoxes. But they can be avoided by incorporating restrictions into the truth theory for a language.[15] Such restrictions may require belief ascriptions to be construed as metalinguistic. And the truth conditions on certain content sentences will be more complicated than my examples suggest. Just what complexities are involved depends on which truth theory is used to interpret the language in which the ascription occurs.

The truth condition yields the virtues of the transparent reading; the classification, those of the opaque. The features required for 'belief' to perform its epistemological function follow from the truth condition. A belief is true just in case its truth condition is satisfied; it is warranted to the extent that there is evidence that its truth condition is satisfied; and so on. The truth condition is what relates (A) to kangaroos and carnivores in such a way that the truth value of the belief ascribed to Sam depends on those animals, and its warrant on what is known about them.

Moreover, the truth condition enables us to explain what is involved in harboring incompatible beliefs and why incompatibilities among beliefs often go unnoticed. Beliefs are incompatible if their truth conditions are not simultaneously satisfiable. Since the agent need not know the truth conditions on his beliefs, he may be in no position to tell whether this is the case. Indeed, sentences that are alike in truth conditions are often classified as disparate *p*-sentences, and on the basis of criteria that are more restrictive than coextensiveness of corresponding terms. As a result, the logical relations among the several sentences an agent cognitively favors are apt to be obscured by their classification. So even if the agent is aware that he cognitively favors certain sentences, he may have no way to tell whether they conflict.

---

[15] Cf. Alfred Tarski, "The Concept of Truth in Formalized Languages," *Logic, Semantics and Metamathematics* (Indianapolis: Hackett, 1983), pp. 152–278; and Saul Kripke, "Outline of a Theory of Truth," this JOURNAL, LXXII, 19 (Nov. 6, 1975): 690–716.

The sentences we classify as instances of a common *p*-sentence are paraphrases of one another. Thus, those which paraphrase 'Kangaroos are carnivores' are all kangaroos-are-carnivores-sentences. The difficulty is to settle on a criterion for paraphrase.

It is tempting, perhaps, to construe paraphrase in terms of synonymy. Then sentences are paraphrases of one another just in case their corresponding terms are synonymous. But we have no clear criterion for synonymy. It is not obvious whether, e.g., 'meat eater' is (close enough to) synonymous with 'carnivore' for the content of (D) to count as a paraphrase of that of (A). Moreover, there are cases in which likeness of meaning, however construed, plays little part in determining the extension of our *p*-sentences. However it is that we decide whether the content of (F) is a paraphrase of that of (E), and so a Lawless-has-been-arrested-sentence, it is plainly not on the basis of anything like synonymy of their corresponding terms.

A related problem is seen in Saul Kripke's puzzle about belief.[16] Pierre, a monolingual French child, sincerely asserts

(J) Londres est jolie.

on the basis, perhaps, of stories he's heard and pictures he's seen. During a war, he is evacuated to a particularly dismal section of London. Since his new acquaintances speak no French, he learns to speak English directly, as a native speaker does. He comes then to assert

(K) London is not pretty.

But being unaware that 'Londres' and 'London' denote the same city, he does not withdraw his previous assertion of (J). What beliefs can we reasonably ascribe to Pierre?

Plainly the beliefs ascribed in

(L) Pierre believes that London is pretty.

and

(M) Pierre croit que Londres est jolie.

have the same truth condition. So the question whether they ascribe the same belief turns on whether they have the same classification condition. Typically we treat translations of p as p-sentences. So ordinarily we have no qualms about classifying (J) as a London-is-pretty-sentence, nor about ascribing to someone who cognitively favors (J) the belief that London is pretty. What gives pause in Pierre's case is the fact that the correct literal translation of a sentence he asserts is one he explicitly disavows.

---

[16] "A Puzzle about Belief," in A. Margalit, ed., *Meaning and Use* (Boston: Reidel, 1979), pp. 239–283.

Sentences, like other entities, can be variously classified; and in different contexts or relative to different interests, different classifications may be appropriate. Frequently we ascribe beliefs in order to explain a person's behavior (including his verbal behavior). If such is our goal here, then we should not treat the translation of (J) into English as a *J*-sentence.[17] For a plausible explanation of Pierre's untroubled assertion of sentences that are incompatible is that he is unaware of their incompatibility. This suggests that we explicate his beliefs as follows:

(M′) $(\exists s)$(Pierre cognitively favors s & s is a Londres-est-jolie-sentence & (s is true $\supset (\exists x)(x =$ London & x is pretty)))

and

(N′) $(\exists t)$(Pierre cognitively favors t & t is a London-is-not-pretty-sentence & (t is true $\supset (\exists y)(y =$ London & $\sim y$ is pretty)))

Pierre's assertion 'Londres est jolie' then is not classified as a London-is-pretty-sentence, despite its being a translation of and having the same truth condition as an English sentence that is so classified. Pierre's beliefs are, on this reading, jointly untenable, but not logically contradictory.

Does this mean that we should entirely abandon the practice of treating translations of p as p-sentences? If we require a general criterion that determines for every p what counts as a p-sentence, we seem forced to do so. This outcome is unwelcome, however; for the practice is useful and rarely gives rise to difficulties. Moreover, without it we have no way to ascribe beliefs that we can understand to people who do not speak our language.

Instead, it seems best to deny that we either have or need such a criterion. p-sentences are paraphrases of p, and there is no general criterion for determining the limits on permissible paraphrase. The goal of an explication of belief ascriptions is, then, not the identification of a single, rigid standard that determines for every R and for every p whether R believes that p.

---

[17] This is not the only reasonable goal. We could take Pierre's situation to show what problems can arise from learning more than one language "like a native." We might then contend that because Pierre learned the languages independently of each other he believes both that London is pretty and that London is not pretty. We would then treat (J) as a London-is-pretty-sentence. Indeed, whatever philosophers think of Kripke's puzzle, it should give the Berlitz people pause.

The quest for such a standard was, in any case, misguided. We seek to explicate actual belief ascriptions, and we are fairly flexible about the beliefs we are prepared to ascribe. We adjust the parameters on permissible paraphrase to suit our several purposes, to answer to our various interests, to achieve our assorted ends. An explication of belief ascriptions should yield a schema that shows what such adjustments involve. It need not yield an algorithm for making them.

It might be thought that I have given up too quickly. Perhaps there are general rules for classifying the sentences we cognitively favor, but more complicated rules than I suggested. If so, the problems I mentioned might just show that the rules have to contain defeasibility conditions or exceptive clauses. For example, all translations of p are p-sentences except when . . . . This suggestion is not promising, however; for the cases that perplex us typically involve mistaken beliefs. And the human capacity to make errors seems far to outstrip our capacity to anticipate them and to build into our rules of interpretation mechanisms to accommodate them as they arise. Unless we can tell in advance just what mistakes will be made, we cannot tell what defeasibility conditions or exceptive clauses need to be incorporated into our rules.

Typically, of course, we have little difficulty deciding how broadly to construe paraphrase, and in what respects paraphrases must agree. So typically we have little difficulty saying which sentences count as paraphrases of each other. But in different contexts and relative to different interests such matters are decided differently.

When we are concerned exclusively with the situation a belief is about, we often take the truth condition to determine the classification condition. Thus, a counterespionage agent interested in getting the goods on Ortcutt might take any sentence *s* whose truth condition is

$$(s \text{ is true} \supset (\exists x)(x = \text{Ortcutt} \,\&\, x \text{ is a spy}))$$

to be an Ortcutt-is-a-spy-sentence, and so take any belief ascription with that truth condition to ascribe the belief that Ortcutt is a spy. This amounts to construing the ascription transparently.

At the opposite extreme, if a sincere avowal is sufficiently bizarre, we might permit nothing but its syntactic replicas to count as paraphrases. If Arlo avers, "People are prime numbers," we might take only replicas of his remark to satisfy the classification condition. In such cases, we are primarily interested in the way a peculiar assertion is to be classified, not in what it would take for that assertion to be true. And, because of its peculiarity, only replicas are counted as satisfying the classification condition. But even though we focus on the classification condition, it is important to notice that these belief ascriptions have truth conditions as well. Indeed, it is because we recognize that their truth conditions are those of blatant falsehoods or trivial truths that we retreat to what amounts to an opaque construal.

Ascriptions of belief typically fall between these extremes. They are, in my terminology, translucent. Sensitivity to contextual factors is thus required to identify the appropriate *p*-sentence and to fix its extension. In the explication of a casual conversation among neighbors, Sam's assertion, 'Kangaroos are carnivores', might be classified as a kangaroos-are-meat-eaters-sentence on the grounds that the coextensiveness of 'carnivores' and 'meat eaters' is widely known. In such a context, however, we should be reluctant to classify it as a marsupials-of-the-family-macripodidae-are-meat-eater-sentence, for the coextensiveness on which that classification is based is not common knowledge. In an explication of a discussion among zoologists, however, his remark might well be so classified. Whether Sam's utterance counts in favor of (B), of (D), of neither, or of both, depends on how it is classified, and its classification depends in turn on a variety of factors including the context in which and the reasons why his assertion is being considered.

The criterion we employ need not take Sam as the final authority. Neither the "common knowledge" that grounds the first classification nor the more specialized knowledge that grounds the second need be knowledge that Sam shares. For sometimes it is reasonable to ascribe to a person beliefs that he does not recognize as his own.

There are, to be sure, cases in which the paraphrases that the agent accepts determine the extension of the *p*-sentence under which they are to be classified. Our basis for classifying the contents of (E) and (F) as Lawless-has-been-arrested-sentences is Max's conviction that they all refer to Lawless. And our reason for refusing to so classify (G) is his conviction that the Sixth Avenue sapphire snatch was perpetrated by someone else entirely. So even though (E) and (G) have the same truth condition, where Max's views on the matter are authoritative, they have different classification conditions. 'Lawless-has-been-arrested-sentence' then does not comprehend every sentence that asserts that Lawless has been arrested. Nor need it comprehend only such sentences. If the Brooklyn bank robberies were in fact perpetrated by someone else, Lawless's capture leaves the Brooklyn bank robber at large. Still, by our criterion, 'The Brooklyn bank robber has been arrested' is a Lawless-has-been-arrested-sentence. (F) then has the same classification condition as (E), but a different truth condition.

These features are not peculiar to cases in which it is the agent's convictions that determine the classification conditions. Sometimes, as we saw, our classification criterion is "common knowledge"—an amalgam of knowledge, ignorance, and error which is publicly shared. Because such "knowledge" falls short of omniscience, there are sentences asserting *p* that are not *p*-sentences. And because it falls short of infallibility, there are *p*-sentences that do not assert that *p*.

The foregoing account yields no criterion of cognitively favoring. As a result, it provides no basis for ruling that some sentences are a priori unbelievable. There may, of course, be sentences that nobody in fact cognitively favors (e.g., sentences with seventeen nested negations). But this is a matter for empirical psychology to discover; it is not to be decided in advance. Consequently, we cannot claim that certain things (e.g., contradictions, violations of natural law, sufficiently improbable events) are intrinsically incredible. This strikes me as a welcome result. For if sentences asserting such things are in fact accepted—if, e.g., they are sincerely uttered, inscribed, or assented to with every evidence of being comprehended—it is better to construe them as irrational beliefs than to deny that they are beliefs at all.

It follows that certain seemingly legitimate inferences cannot be drawn. If *R* believes that *p*, and *R* believes that *q*, it does not follow that *R* believes that *p & q*. For he may be prepared to utter sincerely, inscribe, or assent to each conjunct separately, without being prepared to sincerely utter, inscribe, or assent to their conjunction. This enables us to make sense of harboring beliefs that contradict one another without thereby harboring contradictory beliefs. *R* may cognitively favor *p* and cognitively favor ~*p* without cognitively favoring *p & ~p*. Indeed, doing so is not uncommon. We are all too often willing to assent to each of a pair of contradictory sentences ('Vitamin C prevents colds', 'Vitamin C does not prevent colds') until they are conjoined and a contradiction emerges. Then, since we are unprepared to assent to a contradiction, we revise some of our earlier views.

Perhaps accepting *p* and accepting *q* is prima facie evidence for accepting *p & q*, and hence, other things being equal, for believing *p & q*. But such evidence can be overridden by, e.g., the agent's refusal to utter, inscribe, or assent to *p & q*. So a conjunction principle is not part of the "logic of belief." It is just not true that we always believe the conjunction of the individual statements that we believe.

Our practice in ascribing beliefs is flexible, and the explication of belief ascriptions should do justice to that flexibility. This is something that standard accounts are unable to do. They hope to discover a single rule or criterion in mind which determines the limits on permissible paraphrase in all correct ascriptions of belief. I doubt that there is any such rule. For the limits on paraphrase vary with variations in circumstances which such explications purposely ignore. Any single rule is apt to be too liberal to accommodate some cases or too restrictive to accommodate others.

There is a family resemblance among our ascriptions of belief, but the family is an extended one with any number of odd and embarrassing relatives. Most of the difficulties in the ascription of belief concern mistaken beliefs of one sort or another. And it is unlikely that we will be able to design or discover a single rule that anticipates and accommodates all the ways beliefs can go wrong. Instead of such a rule, I have proposed a schema that enables us to say how individual belief ascriptions function. It recognizes that such ascriptions differ in translucence, so that there is no saying in general whether in paraphrasing a belief ascription we ascribe the same belief or a different one. We have a choice as to how broadly or narrowly to construe the classification condition. And in different circumstances different choices may be appropriate. But we cannot construe it any way we please. Any selection we make is defeasible. If, relative to our interests, a different classification condition or a greater or smaller range of paraphrases yields a better account, then the classification condition we have chosen must be replaced. And if there is no general rule for telling which accounts are better, that is because of the multiplicity and variety of roles that our ascriptions of belief play in our understanding of ourselves and of each other.

To the controversy in the philosophy of mind over the nature, character, or essence of beliefs, I have contributed nothing. The term 'belief' may, for all I have said, denote a physical state, a functional state, an immaterial mental state, or some as yet unimagined alternative. But whatever the nature of such entities, there remains the question of their classification. And there is no more reason to think that mental entities determine their own classification than there is to think that physical entities do. In different circumstances it might be correct to call a physical object a chair, a Hepplewhite, an antique, or an investment. Likewise, in different circumstances it might be correct to call a single mental state a kangaroos-are-carnivores-sentence, a kangaroos-are-meat-eaters-sentence, or a marsupials-of-the-family-macropodidae-are-meat-eaters-sentence. Whether a cognitively favored sentence is identical with or is the counterpart of a mental entity, we still need to know how it is to be classified and under what circumstances it is true. It is the ways we go about answering these questions that my schema is designed to illuminate.

CATHERINE Z. ELGIN
University of North Carolina/Chapel Hill

COMMENTS AND CRITICISM

LEVI'S "THE WRONG BOX"[*]

RECENTLY a number of philosophers have developed decision theories that, in one way or another, explicitly incorporate the idea that a decision maker should assess the value of an available course of action in terms of the extent to which (according to the decision maker's beliefs) the act *causes* good outcomes and *prevents* bad ones. These are the "causal decision theories."[1] What has come to be known as "the traditional theory," however, evaluates the merits of the available acts in terms of the extent to which they would constitute *evidence* (to the agent) for good and against bad outcomes.[2] So this is sometimes called "evidential decision theory." According to causal decision theorists, certain decision problems, called "Newcomb problems," are supposed to tell the two kinds of decision theories apart, causal decision theories there giving the correct answers and evidential decision theory the wrong ones. According to these critics, the eviden-

---

[*] I wish to thank Isaac Levi for generous comments on a previous draft of this paper. Also, this paper was written under grants from the American Council of Learned Societies and the Graduate School of the University of Wisconsin-Madison, which I gratefully acknowledge.

[1] See, for example: Allan Gibbard and William Harper, "Counterfactuals and Two Kinds of Expected Utility," in C. A. Hooker, J. J. Leach, and E. F. McClennen, eds., *Foundations and Applications of Decision Theory* (Boston: Reidel, 1978), pp. 123–162, reprinted in Harper, R. Stalnaker, and G. Pearce, eds., *Ifs* (Boston: Reidel, 1981), pp. 153–190; Brian Skyrms, *Causal Necessity* (New Haven, Conn.: Yale, 1980), chapter IIC, on "The Role of Causal Factors in Rational Decision"; and David Lewis, "Causal Decision Theory," *Australasian Journal of Philosophy*, LIX, 1 (March 1981): 5–30. Robert Stalnaker had earlier proposed such a theory (see *Ifs*, *op. cit.*, pp. 15–18 and 151/2), and Isaac Levi had discussed a related proposal in his "Newcomb's Many Problems," *Theory and Decision*, VI, 2 (May 1975): 161–175.

[2] This theory is developed by Richard Jeffrey in his *Logic of Decision* (New York: McGraw-Hill, 1965; 2nd ed., Chicago: University Press, 1983).

0022-362X/85/8202/0091$01.40

tial theory is not sensitive, in the right way, to certain kinds of causal beliefs.

In a recent exchange of discussion pieces on Newcomb's problem and causal decision theory, Isaac Levi has argued that causal decision theory, as it stands, is in deep trouble since it can give conflicting prescriptions in some decision problems,[3] David Lewis has responded by arguing that Levi's argument commits the fallacy of equivocation and overlooks one of the theory's explicitly stated principles,[4] and Levi has responded by charging, among other things, that Lewis has himself, in his criticism, committed the fallacy of equivocation.[5]

Although I believe that evidential decision theory, if fully and properly exploited, will always give the correct answers in the decision problems that are supposed to tell causal and evidential decision theory apart (so evidential decision theory *is* appropriately sensitive to the relevant kinds of causal belief and need not be given up in favor of causal decision theory), I also believe that causal decision theory, if fully and properly exploited, will always give the correct answers—indeed, the same as those given by evidential decision theory: the same answers as those which causal decision theorists intend their decision theories to give.[6] Of course this implies that causal decision theory never gives *conflicting* prescriptions, as Levi argues that it does.

Levi's and Lewis's discussions concern the two decision problems that Levi considered in the first article in the exchange: what Levi calls "the pseudo Newcomb problem," and the original Newcomb problem itself. And Lewis's responses and Levi's replies discuss the two problems separately (for the most part; Levi sometimes makes cross references and comparisons). I shall here consider the exchanges regarding the two problems separately (for the most part). In each case, after a brief account of the exchange to date, I'll argue that causal decision theory *does* give the intended prescription—for much the same reasons that Lewis gives, but accommodating Levi's replies.

[3] "A Note on Newcombmania," this JOURNAL, LXXIX, 6 (June 1982): 337–342; cited as Levi 1982. The version of causal decision theory that Levi specifically discusses is Gibbard and Harper's theory of *U*-maximization, given in "Counterfactuals and Two Kinds of Expected Utility," *op. cit.*, but most of his remarks would seem to apply equally well to the other causal decision theories cited above.

[4] "Levi against *U*-Maximization," this JOURNAL, LXXX, 9 (September 1983): 531–534; cited as Lewis 1983.

[5] "The Wrong Box," *ibid.*: 534–542; cited as Levi 1983.

[6] For the defense of the evidential theory, see my *Rational Decision and Causality* (New York: Cambridge, 1982), chs. 6–8.

I. THE PSEUDO NEWCOMB PROBLEM

Levi argues that his pseudo Newcomb problem is a case in which causal decision theory gives conflicting prescriptions. In this problem, one opaque box each has already been drawn from two distinct urns. In Urn 1, the bad urn, 90 of the boxes contained nothing, and the remaining 10 each contained a million dollars. In Urn 2, the good urn, 90 of the boxes contained a million dollars and the remaining 10 contained nothing. The agent must decide between these two alternatives: receive only the contents of the opaque box drawn from the good urn (act A_1), and receive a thousand dollars (in a transparent box) plus the contents of the opaque box drawn from the bad urn (act A_2). It is clear that the correct act is A_1. And Levi grants, what is clear and uncontroversial, that causal decision theory can be got to prescribe this act. But he also argues that the theory can be got to prescribe A_2: "for the agent has no causal influence whatsoever over whether the opaque box he selects has a million in it or not. That was 'fixed and determined' well before he makes his decision, just as in the original Newcomb problem" (Levi 1982, p. 340). According to causal decision theories, if a state is outside the influence of an agent's actions, then its probability should enter into the calculation of expected utility unconditional on the act.[7] Presumably, Levi here envisions the causal calculation of the expected utilities of the acts as proceeding as follows, where 'M' denotes the proposition that the million dollars is in the opaque box selected:

$$\begin{aligned} \mathrm{U}(A_1) &= \Pr(M)\mathrm{Des}(A_1 \,\&\, M) + \Pr(\sim M)\mathrm{Des}(A_1 \,\&\, \sim M) \\ &= \Pr(M)\mathrm{Des}(\text{getting } \$1{,}000{,}000) + \Pr(\sim M)\mathrm{Des}(\text{getting } \$0) \end{aligned}$$

and

$$\begin{aligned} \mathrm{U}(A_2) &= \Pr(M)\mathrm{Des}(A_2 \,\&\, M) + \Pr(\sim M)\mathrm{Des}(A_2 \,\&\, \sim M) \\ &= \Pr(M)\mathrm{Des}(\text{getting } \$1{,}001{,}000) + \Pr(\sim M)\mathrm{Des}(\text{getting } \$1{,}000). \end{aligned}$$

It is clear that no matter what the value of $\Pr(M)$ is, the second sum is greater than the first.

In his response to this argument, Lewis asks us to distinguish:

(1) The opaque box the agent selects is such that: the agent has no causal influence over whether or not it has a million in it.

(2) The agent has no causal influence over whether or not: the opaque box he selects has a million in it (Lewis 1983, p. 532).

---

[7] This is the central idea of Skyrms's theory of K-expectation (*op. cit.*). But Gibbard and Harper's theory (*op. cit.*), the causal decision theory Levi specifically discusses, has this as a consequence, as they (and Levi), point out.

Lewis grants that (1) is true. But he claims that (2) is probably false: "It is false, if, for instance, the box drawn from the good urn has a million in it and the box from the bad urn is empty" (Lewis 1983, p. 532). According to Lewis's reasoning, (2) is false if the two boxes have different contents, which has a probability of .82 if the draws from the two urns are independent of each other. Evidently, Lewis's point is that it is (2) that expresses the relevant kind of causal independence of states from acts (or that correctly identifies the relevant state), the kind of causal independence that, according to causal decision theory, requires the probability of the state to be taken to be unconditional on the act in the calculation of expected utility. But, according to Lewis, (2) is (probably) false, and Levi has seized upon the truth of (1)—which does *not* express the relevant kind of causal independence (or which misidentifies the relevant state)—to get the probabilities of the states to be unconditional on the act in the causal calculation of expected utility, thus committing the fallacy of equivocation.

Now for Levi's reply to this. The reply is in basically two parts. First, Levi asks us to consider (1′) and (2′), which read the same as (1) and (2) (minus the colons) but are about the original Newcomb problem. (In this problem, there is only one opaque box, which a highly accurate predictor has already filled with a million dollars if he predicted that the agent will opt for the act of taking only that box (act B_1) and has left empty if he predicted that the agent will take both the opaque box and a transparent box which can be seen to contain one thousand dollars (act B_2), these being the only two acts available.[8]) Levi considers the case in which the agent fully believes that the predictor has predicted his act correctly: "Given this case. . . , if the agent were to choose both boxes, the opaque box selected would contain nothing. If he were to choose the opaque box alone, the opaque box selected would contain $1,000,000. Thus, utilizing the reasoning that seems to get Lewis to the conclusion that (2) is probably false, I conclude that (2′) is probably false as well" (Levi 1983, p. 535). And if (2′) is indeed false, causal decision theory will prescribe the one-box act in the original Newcomb problem, the act that causal decision theory was designed to avoid.

---

[8] Causal decision theorists have argued that this decision problem tells evidential and causal decision theories apart, the evidential theory giving the one-box act and the causal theory prescribing two boxes. Levi, in "Newcomb's Many Problems" (*op. cit.* 1975), argues, however, that the evidential theory will give the one-box act only if Pr(predictor predicts B_i/B_i) is high for both $i = 1$ and $i = 2$, which, he argues, doesn't follow from the standard descriptions of Newcomb's problem. That this is not so, however, is argued in my "Newcomb's Many Solutions," *Theory and Decision*, XVI, 1 (January 1984): 59–105. What is required is simply that B_i be positively probabilistically relevant to the prediction that B_i, at least to some (small) degree.

Of course the conditionals in Levi's argument must be construed as "backtracking" conditionals if they are to be true. That is, the conditionals can't both be true if the predictor's (prior) prediction is held fixed across the evaluation of the truth values of the two conditionals. So, the only way in which it could follow from the truth of these two conditionals that (2′) is false (i.e., that the agent *does* have *causal* influence on the contents of the opaque box) is if this causal influence goes backwards in time. So if causal influence does *not* go backwards in time, (2′) would seem to be true, and so causal decision theory will give the two-box solution. (More to the point, and more accurately: if the agent does not *believe* in backwards causation, then he will *believe* (2′) to be true, in which case causal decision theory will give *him* the two-box solution.)

In the case of (2), in connection with the pseudo Newcomb problem, however, its *falsity* is consistent with the assumption that there is no backwards causality (and, of course, with the assumption that there is). Assume, what is highly probable, that the box from the good urn contains a million and that the box from the bad urn contains nothing. Then the agent must make his choice *before* it is "fixed and determined" that the chosen box has a million dollars (or nothing) in it; that is to say: *not until the agent has made his choice will it be the case that he has chosen a box with (or without) a million dollars in it.* This is *not* to say, of course, that the agent has causal influence over the contents of the particular box he happens to choose, but only that he has causal influence over whether or not the box he chooses *is the one with the million.* (This is clearly the relevant state to ask about the agent's influence upon, as elaborated on below.)

In the second part of Levi's reply, although he says that he did not equivocate between (1) and (2), (2) being his intention throughout, he admits that he may have equivocated between the following:

(2.1) The agent has no causal influence over whether or not the opaque box he selects has a million in it through his choosing the contents of the transparent box and the box from urn 1 or the contents of the box from urn 2.

(2.2) The agent has no causal influence over whether or not the opaque box he selects has a million in it through his choosing to receive the contents of the transparent box and the opaque box from urn 1 when the box from urn 1 is empty and the box

> from urn 2 has a million, or his choosing to receive the contents of the box from urn 2 when the box from urn 1 is empty and the box from urn 2 has a million (Levi 1983, p. 536).

Levi concedes that (2.2) is false, but insists that (2.1) is true. This is argued on the basis of a seemingly plausible view about the relation between causation (between propositions) and explanation: "no matter how we proceed to further analyze such causal dependency or the conditions for causal explanation, such causal dependency presupposes causal explainability" (Levi 1983, pp. 536/7). As to (2.2), Levi says it is false, because "one can explain why the box selected by the agent contained nothing (a million) by asserting that he chose the opaque box from urn 1 (urn 2) when the box from urn 1 contained nothing and the box from urn 2 contained a million" (Levi 1983, p. 537). In any case, it is, of course, Levi's argument in connection with (2.1) that is more important, for it is this that he argues is true, and which if true would leave causal decision theory prescribing the one-box act, conflicting with its prescription of two boxes.

Levi says:

> I claim that (2.1) is true *even when the box from urn 1 contains nothing and the box from urn 2 contains a million*. If (2.1) were false, I could explain why it is true that the opaque box contains nothing (one million) by asserting that he chose to receive the contents of the opaque box from urn 1 (urn 2). But that is no causal explanation even under the circumstances envisaged. To be sure when the box from urn 1 contains nothing and the box from urn 2 contains a million, one can offer a (true) explanation by asserting that the agent chose to receive the contents of the opaque box from urn 1 (urn 2) when the box from urn 1 contains nothing and that from urn 2 contains a million. Moreover, the explanation is noncontroversially causal. But such explanation simply is not forthcoming through appeal to whether the agent chose the transparent box and the box from urn 1 or the box from urn 2—no matter what the contents of the two opaque boxes happen to be (1983, p. 536).

There are two things wrong with this argument: first, the conclusion [that (2.1) is true] is probably false; and second, the version of the principle of causation's presupposing explanation which Levi needs to get that conclusion is certainly false.

Let's call the box drawn from urn 1 "Box 1" and the one from urn 2 "Box 2." Clearly the agent doesn't have any causal influence over the contents of either of these boxes. Suppose the agent selects Box 1 (together with the transparent box), and so Box 1 is "the selected box." Clearly the agent doesn't have any causal influence over the contents of the selected box in the sense of having causal influence over the contents of Box 1. And supposing that Box 2 is "the selected box," the agent again clearly doesn't have causal influence over the contents of the selected box in the sense of having causal influence over the contents of Box 2. But of course this can't be what Levi means when he denies that the agent has causal influence over the contents of the selected box. For this is only to grant the truth of Lewis's (1). And Levi wants to argue also for the truth of (2), construed of course as (2.1).

Of course the agent doesn't care about the boxes qua boxes. So let's talk money instead of "selected box." What's relevant here, obviously, is how much of it he receives, and how the acts causally influence how much of it he receives.[9] Suppose again, what is very probable (probability of .81), that Box 1 contains nothing and that Box 2 contains a million. Given the truth of this, it seems clear that, through choosing Box 1 and the transparent box, the agent would cause himself to receive $1,000, and that, through choosing Box 2, he would cause himself to receive $1,000,000. If the contents of the boxes were reversed (probability of .01), then the two acts would have quite opposite causal effects. And if the two boxes contained the same amount of money (probability of .09 + .09 = .18), then the acts would not make much difference as to how much the agent receives. Of course there is no single causal role of a given act: in one constellation of causally relevant background factors (i.e., distribution of money between the two boxes), choosing Box 2, for example, would cause the agent to become a millionaire, and in another constellation, it wouldn't. But of course this doesn't diminish the agent's *probable* causal influence over what his winnings will be.

But what about Levi's reductio of the falsity of (2.1), using his idea about causation's presupposing explanation?

Levi's argument for the truth of (2.1) is that, if it were false, then "I could explain why it is true that the opaque box contains nothing (one million) by asserting that he chose to receive the contents of the opaque box from urn 1 (urn 2)." Of course the explanation is not supposed to be of the fact that *Box 1* contained such and such or that *Box 2* contained such and such, but rather of the fact that the *opaque box selected* contained such and such, in the sense that *the agent will walk to the bank with such and such*. But clearly we *can* explain this by reference to the act: we just also have to throw into the explanation (into the explanans) the information about what amounts of money were in the two boxes. But perhaps Levi's point, then, is that if (2.1) were false, then reference to the act *alone* would be sufficient to explain why the agent became (or didn't become) a millionaire, nothing else being explanatorily relevant—where clearly reference to the act *alone* does *not* fully explain this. Then it would seem that Levi's view about the relation between causation and explanation would have to imply something like this (which I'll call "Proposition *"): If X would cause Y and $\sim X$ would cause $\sim Y$, then a Y (a $\sim Y$) could be explained just by reference to X (to $\sim X$) (and perhaps also using other things that we know).

---

[9] The same points could be made below, talking boxes instead of money. But this will make it easier to avoid the ambiguity of "selected box" which Lewis points out.

What sorts of things are X and Y above? Recall that Levi's views about the relation between causation and explanation has to do with causation between *propositions*. His view is that *such* causal dependency doesn't make sense unless it presupposes explainability. Evidently, the reason for Levi's talking of causation between propositions here—instead of between *events*—is that it seems that this is the way in which most causal decision theorists think of causation. In any case, there are two options open to causal decision theorists here in defense of the claim that (2.1) is probably false. First, they could say that their talk of causal independence of outcomes from acts really has to do with the *events* of the relevant acts' being performed and of the relevant outcomes accruing to the agent [in which case Proposition * is clearly false and (2.1) is clearly probably false, as explained below], *or*, second, they could construe their talk of causal independence of outcomes from acts in terms of causal dependency between propositions, and then explain, e.g., as below, how there can nevertheless be causation without explainability, as in the case of the probable falsity of (2.1).

Proposition * is clearly false if X and Y are construed as events. Suppose it is true that your typewriter is on its last legs, but that you don't know this. Then pressing the carriage return might *cause* the coils in its motor to overheat and smoke. But of course you can't *explain* the smoking of the motor just by asserting that the carriage return was pressed. If you let an old car idle for too long, the engine can heat up. Then stepping on the accelerator will either cool the engine down by circulating the coolant, or cause it to heat up even more if there is no coolant (or if it is blocked, etc.). So stepping on the accelerator can *cause* the car to overheat without its being possible to *explain* the overheating by asserting simply that the accelerator was depressed: you also need to specify the state of the cooling system, just what was wrong with it.

That *X* would cause *Y* in a given situation does not imply that it would cause *Y* in every situation. As I understand it, this is (part of) the point of J. L. Mackie's idea of a "causal field."[10] In one causal field (e.g., cooling system in good condition) an event *X* (e.g., depressing the accelerator) would cause *Y* (e.g., motor cools down), but in another (e.g., out of coolant), *X* wouldn't cause *Y*. In the pseudo Newcomb problem, the causal fields are the four different possible distributions of money between the two boxes. And to get an *explanation* of *Y*, you need both the causal field and reference to an event *X* which occurred, and which, in that causal field, causes *Y*.

And of course the explanation for why there can be causation without explanation is that causation between events is a relation between the cause, the effect, and the causal field, whereas an additional relatum in the relation of explanation is the person who gets explained to, who may not be aware of the operative causal field. In the most probable case of the pseudo Newcomb problem, the event of taking just the one opaque box would cause the event of your becoming a millionaire, and the event of taking two boxes would cause you to receive only a thousand. But reference to the acts *alone* wouldn't *explain* why their effects occurred: to get an *explanation*, you also need to know the contents of the relevant box.

Of course if one wanted one's causal talk to be understood in such a way that "*X* causes *Y*" implies that *X* alone is causally sufficient for *Y*, then (2.1) just wouldn't express the (clearly intended) condition that must be satisfied for the probabilities of the outcomes to be taken to be unconditional on the act in the causal calculation of expected utility. It seems clear that if causal decision theorists wanted their causal talk to be understood in this way, then that condition would have to be expressed in terms of whether or not the acts would *complete* causally sufficient (or probabilistic) conditions for the various outcomes.

Let's now construe '*X*' and '*Y*' in Proposition * as ranging over propositions. Presumably, the controlling idea here is that whether or not a particular causal relation obtains between two propositions is dependent on how the relevant events are *described* in the propositions. Thus, its being true that the accelerator is depressed does not cause it to be true that the engine overheats. But its being true that the accelerator is depressed *when there is no coolant* does cause it to be true that the engine overheats. In any case, the causal decision theorist's response is clear. As indicated in the previous paragraph, if one understands causal dependency in such a way that a cause must alone be a causally sufficient (or probabilistic) condition for its effect, then the causal decision theorist simply shifts from talk about "causes" to talk of *completions* of causally sufficient (or probabilistic) conditions. Thus, understanding causal dependency as a relation between propositions, the question is whether or not the truth of a proposition to the effect that a particular act is performed would *complete* a causally sufficient (or probabilistic) condition for the truth of a proposition that a particular outcome will accrue to the agent—and of course whether or not it would complete such a condition will depend on what *other* propositions happen to be true. Again, we can have causation without explainability of the effect merely by reference to the cause, because it may not be known what other propositions are true.

---

[10] See *The Cement of the Universe* (New York: Oxford, 1980). This idea of a causal field bears important affinities, incidentally, to Lewis's "dependency hypotheses" ("Causal Decision Theory," *op. cit.*) and to Skyrms's propositions "H_n & K_{ni}" (*Causal Necessity, op. cit.*).

So it seems clear that (2.1), properly understood, is very probably false, and that the truth is: The probability is .81 that A_1 would cause the agent to receive a million, and A_2 would cause him to receive only a thousand; the probability is .01 that A_1 would cause the agent to receive nothing, and A_2 would cause him to receive a million and a thousand; the probability is .09 that either act would make the agent a millionaire; and the probability is .09 that neither act would make him a millionaire. Straightforward calculation (indeed, see Levi 1982, p. 340) shows that causal decision theory gives the right answer: A_1, i.e., take just the contents of Box 2.

II. THE ORIGINAL NEWCOMB PROBLEM

Levi also argues that, in the original Newcomb problem, causal decision theory delivers conflicting prescriptions. In this problem, recall, there is only one opaque box, which a very accurate predictor filled with a million dollars if he predicted that the agent would take only that box (act B_1) and left empty if he predicted that the agent would take both boxes (act B_2, where these are the agent's only two options). Levi grants, what is clear and uncontroversial, that the theory can be got to prescribe the two-box act. But he argues that it can also be got to give the one-box answer: "To obtain this result, I identified the 'possible outcomes' of the options involved differently than Newcombites are accustomed to do [namely, with the acts themselves, which surely are within the agent's influences]; but since nothing in their formal principles or informal arguments prohibit this maneuver, it is entirely fair to conclude that pending further elucidation the theory of U-maximization is inconsistent" (Levi 1982, p. 341). (In both causal and evidential decision theory, the desirability of an outcome is equal to its evidential expected utility.)

And Lewis's response to this argument is simply that there is a principle of causal decision theory that prohibits taking these acts to be outcomes, and Lewis quotes Gibbard and Harper's formulation of the principle: "An outcome is complete in the sense that any further specification of detail is irrelevant to the agent's concerns."[11] The point is that each act in Newcomb's problem is consistent with the agent's getting a million dollars and with his not getting it; and Lewis assumes that this is relevant to the agent's concerns.

Levi's reply,[12] in part:

> We shall have to ask Gibbard and Harper what they intended when they wrote these remarks and whether they put them in to cover problems like the one under consideration. I confess that teasing Lewis's interpretation out of their remarks in the context in which they appear transcends my hermeneutical abilities (1983, p. 539/40).

But consider the act B_1, for example. Clearly there is a further specification of what might happen which is relevant to the agent's concerns. We can add more detail to B_1 and make it better or worse. Adding the detail that the agent gets a million dollars makes it better, and adding the detail that the agent won't get a million makes it worse. Gibbard and Harper say that an outcome of an act is a "proposition which, for all the agent knows, expresses *all* the [possible] consequences of that act which he cares about" (*op. cit.*, sec. 3). Thus, assuming that the agent cares about receiving or not receiving a million dollars, the outcomes should specify whether or not the agent receives a million dollars. Levi points out, however, that the agent will probably not be indifferent between the various consequences of receiving a million dollars. So perhaps it should be insisted that satisfaction of Gibbard and Harper's definition of "outcome" requires specification of these consequences, which will in turn have consequences of their own, and so on.

I agree that it is probably impossible to specify an outcome in the idealized way given by various authors of decision theories. F. P. Ramsey describes outcomes as "the different *totalities* of events between which our subject chooses—the ultimate organic unities."[13] And Leonard Savage describes them as "what the person experiences," remarking later that "this idea of pure experience, as a good or ill, seems philosophically suspect and is certainly impractical." And he goes on to point out that Richard Jeffrey's theory (now known as the "evidential theory") affords "escape from this suspect concept."[14] This is because the evidential expected utility of an item is the same no matter how the space of all possibilities is carved up into outcomes and states.

---

[11] 1983, p. 534, quoted from Gibbard and Harper, *op. cit.*, sec. 3.

[12] This is actually only the first part of Levi's two-part reply, but the second part needn't detain us, for it relies on the truth of (2.1), which, I have argued (following Lewis, I believe), is probably false.

I have argued elsewhere[15] that this last consideration constitutes a significant advantage of the evidential theory over certain versions of causal decision theory; for some versions of the causal theory (variations on Brian Skyrms' *K*-expectation theory) will give different values for the acts given different partitions of states. Levi also points out this fact, remarking also that varying the partition of outcomes has the same effect for the theory of *U*-maximization (1982, p. 342, and 1983, p. 541). But what I argued was that the problem for causal decision theory here had to do with coming up with appropriate partitions into *states* of two different *kinds*. The problem for the versions of causal decision theory I was considering had to do with distinguishing two different kinds of states (which two kinds can be very roughly, but inadequately (I argued), described as "those within" and "those outside of our causal influence"), and not with distinguishing states from outcomes. As explained below, I don't believe there is a serious problem for causal decision theory in connection with the distinction between outcomes and factors of other kinds.

As to Levi's point about the outcomes, there seems to be a rather natural way of clearly formulating the intentions of causal decision theorists in this regard. The kinds of considerations which motivated the development of causal decision theory had to do not with the descriptions of outcomes, but rather with importantly different ways in which acts can be *probabilistically relevant to decision makers' happiness*. And this would seem to be a consideration completely unrelated to the problem of *outcome description*. Recall also that the values of outcomes are not calculated; they are assumed and then used in the calculation (causal or evidential) of the values of other, less specific, items. So, for the purpose of assessing the adequacy of the causal *revision* of traditional Bayesian decision theory, it seems fair to suppose that a factor is an outcome only if it makes no difference whether its value is assessed on the evidential paradigm or on the causal paradigm. That is: a factor is an outcome only if it is sufficiently detailed that its evidential expectation equals its causal expectation for every way of partitioning the outcome itself into further outcomes of its own. This suggestion has the advantage that an outcome would specify *enough* of what the agent cares about for causal decision theory to give a unique prescription, without having to specify *literally everything* the agent cares about. And, on this understanding of "outcome," the acts in Newcomb's problem would *not* count as outcomes, but specifications of the agent's possible winnings *would* (assuming, of course, that all the good or bad associated with the winnings are believed to be causal *effects* of winning the relevant amount).

---

[13] "Truth and Probability," in Henry E. Kyburg and Howard E. Smokler, eds., *Studies in Subjective Probability* (New York: Wiley, 1964): 61–92, p. 77 (emphasis mine).

[14] "Difficulties in the Theory of Personal Probability," *Philosophy of Science*, XXXIV, 2 (June 1967): 305–310, p. 306.

[15] *Rational Decision and Causality*, *op. cit.*, ch. 5. There I suggest "appropriateness conditions" for the relevant partitions.

In any case, Levi also charges Lewis's response with equivocation, in connection with Lewis's different treatments of the pseudo and original Newcomb problems. Levi asks us to consider the following two propositions about the original Newcomb problem:

> (2.1′) The agent has no causal influence over whether or not the opaque box he selects has a million in it through his choosing the contents of the transparent box and the opaque box or the contents of the opaque box alone.
>
> (2.2′) The agent has no causal influence over whether or not the opaque box he selects has a million in it through his choosing the contents of the transparent box and the opaque box when the demon predicts correctly, or through his choosing the contents of the opaque box alone when the demon predicts correctly (1983, p. 537).

As with (2.1), Levi claims that (2.1′) is true, since "one cannot causally explain why the opaque box selected contained nothing (a million) by asserting that the agent chose to receive the contents of two boxes (the opaque box alone)" (1983, p. 537). And, as with (2.2), he says that (2.2′) is false, because, he argues, one can in the circumstances described in (2.2′) give the relevant (causal) kind of explanation for the contents of the opaque box. It will fortunately not be necessary to work through the argument offered for this last claim in order to make my main point, below, about this part of the controversy.

Levi charges equivocation as follows:

> Why do [U-maximizers] recommend the one-box option in the pseudo-Newcomb problem on the grounds that (2.2) is probably false, while favoring the two-box option in the Newcomb problem because (2.1′) is true? Or, to put the matter somewhat differently: why do they interpret (2) so that it is probably false, while interpreting (2′) so that it is certainly true? (1983, p. 538)

The answer (presuming to speak for U-maximizers) is simply that this is not what U-maximizers are doing. Rather, U-maximizers interpret (or at least it seems to me they should) (2) as (2.1) and (2′) as (2.1′). Levi agrees that (2.1′) is true. And I have argued that (2.1) is probably false. Hence, interpreting (2) and (2′) in the *same* way (in the *relevant* way) distinguishes the two problems: (2) is probably false and (2′) is true. Hence, causal decision theory prescribes A_1, but B_2.

ELLERY EELLS

University of Wisconsin/Madison

EPICYCLES

IN "A Note on Newcombmania,"[1] I charged that U-maximizing is infected with inconsistency. This allegation was predicated explicitly on the point that the recommendations grounded on U-maximization depend upon how "possible outcomes" of options are described (339).

Obviously it is open to defenders of U-maximization to repair the inconsistency by seeking some way to favor certain types of descriptions of consequences over others. About the time I wrote "A Note on Newcombmania," David Lewis made such a proposal.[2] And there are many others that might be made.

In "The Wrong Box",[3] I did not try to show that the theory of U-maximization as modified by Lewis is inconsistent. To the contrary, I proposed a way of removing an ambiguity in Lewis's discussion and then concluded that U-maximization recommends picking two boxes in both the pseudo Newcomb problem and the

---

[1] This JOURNAL, LXXIX, 6 (June 1982): 337–342.

[2] "Causal Decision Theory," *Australasian Journal of Philosophy*, LIX, 1 (March 1981): 5–30.

[3] This JOURNAL, LXXX, 9 (September 1983): 534–542.

0022-362X/85/8202/0104$00.50

Newcomb problem (540). Lewis can rescue *U*-maximization from inconsistency in this way, but not from absurdity.

Of course, my argument (534/5) rested on another contentious issue. I claim that either both (2) and (2′) are probably false or both are true depending upon how these claims are construed. On this basis, I charged that Lewis equivocated in holding (2) probably false while tacitly maintaining that (2′) is true. If I am right, reading (2) and (2′) so that they are both true is the charitable construal of *U*-maximizing theory with the Lewis emendation. It does not then lead to inconsistency; but it does yield absurdity.

As I wrote in that note:

> I have no doubt that we may expect more epicycles to be added to the edifice of causal decision theory to protect its consistency and to mitigate some of its more obviously absurd recommendations. I suspect we will be invited to accept some conception of causality which will furnish a rationale for taking (2′) to be true and (2) probably false (541).

My expectations are partially fulfilled by Eells' comment, "Levi's 'The Wrong Box'."[4] Eells seeks to explain why it is plausible to think of (2) as probably false while taking (2′) for true (96-100).

For example, when considering causality as a relation between events, he advocates what I do not deny—to wit, that we may cite one event as a cause of another without presupposing that we have provided a causal explanation (98). Indeed, I explicitly endorsed just this Davidsonian point and observed that, in both the Newcomb and pseudo Newcomb problems in the probable case, one cannot causally explain why the agent received a million by mentioning that he chose one box even though one may correctly say that the choosing of the one box is the cause of his receiving a million (*op. cit.*, p. 538, n. 2). Eells seems to use the point to support the probable falsity of (2), but fails to appreciate that if it does sustain the probable falsity of (2), it supports the probable falsity of (2′) as well.

According to the Davidsonian approach I favor, one event (or occurrence) can be the cause of another. But a proposition cannot be the cause of another proposition. Of course, the truth of one proposition may be causally explained by appeal to the truth of other propositions. Hence, if one wants to speak of causal dependency between propositions, such talk ought to be construed as talk of some species of explanation. That is the force of my remarks about causal dependency presupposing explainability. It precludes Eells' tentative interpretation of my remarks as endorsing his condition*.

---

[4] This JOURNAL, this issue, 91-104.

Eells' misunderstanding of my argument seems predicated on a commitment to some kind of conditional analysis of causation of the sort favored by J. L. Mackie (whom he cites), by Lewis, and by many others who oppose the Davidsonian attitude I have adopted. Although I am far from convinced that conditional analyses of causation drive one to the probable falsity of (2.1) and the truth of (2.1′), it is possible that someone might seek to build the right mix of epicycles to yield this result. That is why I wrote the passage about epicycles quoted above. By way of contrast, the Davidsonian approach requires that one treat the Newcomb and pseudo Newcomb problems alike.

Of course, one might complain that I have adopted a controversial view of causality and, to that extent, I have invoked my own metaphysical predilections to beg the question against *U*-maximizers, just as, so I charge, Eells has done against my view.

Not so! My point may be put thus: Whether one endorses the Davidson-like view I favor or some variant of a conditional approach to causation, Bayesian decision theory is applicable. To avoid contradiction or absurdity, *U*-maximization needs to reject the Davidsonian approach. If this is right, Bayesian decision theory has an admirable immunity from the vagaries of disputes in causal metaphysics which *U*-maximization lacks. As I wrote in "The Wrong Box":

> Thus, not only is *V*-maximization robust and relatively free of the epicycles that threaten to overwhelm *U*-maximization, but its own conception of the relevance of causal judgement to choice is relatively straightforward, demands relatively noncontroversial imputations of causal efficacy and commits few hostages to any specific approach to the analysis of causality (541/2).

Eells' discussion illustrates and confirms how important and valuable this feature of *V*-maximization is.

ISAAC LEVI

Columbia University

The trustees and editors of the *Journal of Philosophy* are deeply grieved to report the death of

Albert G. Redpath

President of the Board of Trustees of this JOURNAL from July of 1954 until he died, on Monday, December 17, 1984.

Mr. Redpath's interest in philosophy dates from his days as an undergraduate at Columbia, where he studied under Frederick J. E. Woodbridge, founder of the *Journal of Philosophy*, and where he was a fellow student of three later members of the Columbia Philosophy Department, Horace Friess, James Gutmann, and John Herman Randall, Jr., who preceded him as President of the Journal Board. Mr. Redpath attended Columbia Law School and became a lawyer and investment banker. At the time of his death he was 88 years old.

During his long tenure as trustee of the *Journal of Philosophy* he was generous with his time and energy, his support, and his help in managing its financial affairs. The editors and trustees will miss the benefit of his sharp intelligence and his good company.

NOTES AND NEWS

The Austrian Ludwig Wittgenstein Society is pleased to announce the Tenth International Wittgenstein Symposium, from August 18 (arrival) to August 25, 1985 (departure). in Kirchberg/Wechsel (Lower Austria, near Vienna). The theme of this year's symposium will be "The Tasks of Contemporary Philosophy". The following sections are planned: (1) Wittgenstein; (2) Philosophy as the foundation and methodology of the sciences and humanities; (3) Philosophy as logic and critique of language; (4) Our moral responsibility for human beings and the environment (a) in science and technology, (b) in economics and politics; (5) The problem of a contemporary metaphysics; and (6) Is philosophy a substitute for religion and "Weltanschauung"? A seminar on Austrian Philosophy will be part of the program of the conference.

Scholars wishing to present a paper on a topic that falls within the above themes (papers for sections 2-6 need not be connected with Wittgenstein's philosophy) should request an abstract style sheet from either the President of the Austrian Wittgenstein Society, Adolf Hübner, Markt 234, A-2880 Kirchberg am Wechsel/Austria, or Werner Leinfellner, University of Nebraska, Department of Philosophy, Lincoln 68508/USA. The deadline for receipt of abstracts is June 15, 1985. Should more papers be offered than can be accepted, selection will be by decision of the organizing committee. Contributors will be informed of the committee decision not later than July 15, 1985. Official languages will be English and German.

Those planning to attend the conference should write to Dr. Hübner at the above address for further information as soon as possible. Conference fees will be 1.200 Austrian schillings for nonstudents and 400 Austrian schillings for students.

The University of St. Andrews is pleased to announce an interdisciplinary conference on Psychoanalysis and the Philosophy of Mind, to take place in the University of St. Andrews March 19-21, 1985, at the end of the 1985 Gifford Lecture Series on Psychoanalytic Theory and Science which will be given by Adolf Grünbaum, Andrew Mellon Professor of Philosophy and Research Professor of Psychiatry at the Center for Philosophy of Science, University of Pittsburgh. Invited speakers include: F. Cioffi (Essex), F. Crews (Berkeley), I. Dilman (Swansea), M. Eagle (York, Ontario), J. Haldane (St. Andrews), A. Hobson (Harvard), J. Hopkins (Kings College, London), M. Moore (U.S.C., Los Angeles), R. Sharpe (Lampeter, University of Wales), K. Wilkes (St. Hilda's College, Oxford). Requests for further information should be sent to Peter Clark, Department of Logic & Metaphysics, the University of St. Andrews, Fife, Scotland KY16 9AL.

The dates and titles of Professor Grünbaum's Gifford lectures are as follows: Feb. 12, The Rationale for the Psychoanalytic Conception of Unconscious Motives; Feb. 14, Does *Therapeutic* Evidence Cogently Support the Psychoanalytic Aetiologies? Feb. 19, Are There Any *Freudian* Slips at All? Feb. 21, Are All Dreams Instigated by Repressed Infantile Wishes? Feb. 26, Latent Motives: Reasons but not Causes of Human Conduct? Feb. 28, Rival Perspectives on the Nature of the Psychoanalytic Enterprise; March 5, The Futility of Abandoning Scientific Goals in Psychoanalysis; March 7 and 11, Cognitive Flaws in the Psychoanalytic Method: Part I, Bogus Clinical Data, and Pseudo-consilience of Inductions, and Part II, Epistemic Defects in the Clinical Validation of Causal Hypotheses; March 12, Psychoanalysis and Theism.

On 2-4 November 1984 the American Section of the International Association for Philosophy of Law and Social Philosophy, a group composed mainly of philosophers, lawyers, and political scientists, held its biennial meeting at Notre Dame University, in Indiana. The following resolution was passed: "We deplore the lack of progress towards reduction of nuclear weapons in recent years. We call on all those in a position to do so to make renewed, serious, and sincere efforts to reduce the enormity of the threat to humanity constituted by the present level of nuclear armaments. This includes a call to those who teach to encourage their students and colleagues to think in ways more conducive to survival than now prevail."

The 1985 conference of the Academy for Jewish Philosophy will take place April 28/29, 1985 in Philadelphia. The theme of the conference will be "Judaism and Ethics"; the program will be sponsored by the Religion Department of Temple University together with Congregation Beth Zion—Beth Israel. More information may be obtained from Norbert M. Samuelson, Religion Department, Temple University, Philadelphia, Pa. 19122.

The American Society for Aesthetics invites papers for its forty-third annual meeting at the University of Louisville in Louisville, Kentucky, from October 23 to 26, 1985. Papers are welcome on any topics in the philosophy of the arts and literature, aesthetics and the history of aesthetics, and issues in contemporary criticism. They should not exceed 12 pages and should be accompanied by a one-page abstract. Papers and abstracts (in triplicate) should be sent by April 1, 1985, to Mary Wiseman, Program in Philosophy, The Graduate School of the City University of New York, 33 West 42nd Street, NYC 10036.

The Long Island Philosophical Society is sponsoring a Conference on Ethics and the Environment, April 13, 1985, at the C.W. Post Campus of Long Island University, Greenvale, New York. The conference is intended to address particular environmental issues of major concern—for example, nuclear power development, waste disposal, safe ground water supply—and seeks participation from philosophers as well as from other academic disciplines and professions. Further information may be obtained from Richard E. Hart, Continuing Education, Long Island University, C.W. Post Campus, Greenvale, NY 11548.

The Francis Bacon Library of Claremont, California announces that the 1985 Bacon Birthday Celebration was held on Tuesday, January 22. Sponsored by the Library in cooperation with the Claremont Graduate School, the Bacon Celebration is now in its twelfth year. Ian Box, Mt. Allison University, New Brunswick, Canada, spoke on "Power/Knowledge in Bacon's Natural Philosophy."

The California State University at Fullerton announces its 15th annual philosophy symposium, "Philosophy and Women: The Past 25 Years," which will take place Feb. 27–March 2, 1985. The symposium, which this year is a part of CSUF's 25th Anniversary Celebration, will address normative issues of contemporary interest and will promote discussion between philosophers and nonphilosophers whose professional interests are related to the topic. Requests for more detailed information should be directed to Betty C. Safford, Lecturer in Philosophy and Women's Studies, California State University, Fullerton 92634.

The Department of Philosophy of The University of Chicago announces a conference on Greek ethics to be held April 11/12, 1985. Principal speakers will be John Cooper (Princeton), Terence Irwin (Cornell), Anthony Long (Berkeley), and Nicholas White (Michigan). Further information is obtainable from Ian Mueller, 1050 East 59th Street, Chicago, Ill. 60637.

Christopher Newport College will sponsor a conference on Critical Thinking April 11/12, 1985, under the auspices of the Association for Informal Logic and Critical Thinking and in conjunction with the annual meeting of the Virginia Humanities Association. The goal of the conference is to direct the attention of educators, humanists and scientists to recent work on critical thinking and informal logic in order to help address the problems and needs of American education today. Further information may be obtained from John Hoagland, Department of Philosophy, Christopher Newport College, 50 Shoe Lane, Newport News, Va. 23606.

The Department of Philosophy at Duke University is pleased to announce the appointment of Fred I. Dretske as the Robert Leet Patterson Distinguished Visiting Professor for Spring 1985.

The Department of Philosophy at Duke University and the Duke University Law School will co-sponsor a conference on the philosophical issues of criminal responsibility on April 18 and 19, 1985.

The Minnesota Center for the Philosophy of Science, in conjunction with the Charles Babbage Institute for the History of Computing, has organized a conference on the history and philosophy of modern mathematics, to be held on the campus of the University of Minnesota (Twin Cities) from May 17 to May 19, 1985. Among the speakers will be Richard Askey (Mathematics, Wisconsin), Garrett Birkhoff (Mathematics, Harvard), Felix

Browder (Mathematics, Chicago), Thomas Hawkins (Mathematics, Boston University), Harold Edwards (Mathematics, NYU), Lynn Steen (Mathematics, St. Olaf), Michael Friedman (Philosophy, Illinois/Chicago), Howard Stein (Philosophy, Chicago), Warren Goldfarb (Philosophy, Harvard), Philip Ehrlich (Philosophy, Minnesota), Gregory Moore (History of Science, Toronto), Lorraine Daston (History of Science, Princeton), Michael Crowe (History of Science, Notre Dame), Judith Grabiner (History of Science, California State), Joseph Dauben (History, CUNY), William Aspray (Babbage Institute, Minnesota), and Philip Kitcher (Philosophy, Minnesota). Further information can be obtained from Philip Kitcher, Minnesota Center for the Philosophy of Science, 315 Ford Hall, 224 Church Street S. E., Minneapolis 55455.

The Department of Medical Jurisprudence and Humanities, University of Nebraska Medical Center, is pleased to announce the appointment of Andrew Jameton as Assistant Professor effective October 22, 1984.

Steven M. Cahn, Professor of Philosophy at The Graduate School and University Center of The City University of New York, has been appointed Provost and Vice President for Academic Affairs at that institution.

The Department of Philosophy of the University of North Carolina at Greensboro is pleased to announce that its ninth annual symposium will be held on March 29-31, 1985, focusing this year on ethical theory. The principal speakers will be Lawrence Blum (Massachusetts/Boston), Dan Brock (Brown), Norman Dahl (Minnesota), Stephen Darwall (Michigan), and Patricia Greenspan (Maryland). Commentators include Harvey Green (Tulane), Thomas Hill, Jr. (North Carolina/Chapel Hill), Terrance McConnell (North Carolina/Greensboro), Lance Stell (Davidson), and Laurence Thomas (North Carolina/Chapel Hill). Further information can be obtained from the Department of Philosophy, Foust Hall, University of North Carolina at Greensboro, 27412-5001.

The Department of Philosophy of The Ohio State University is pleased to announce its second Working Conference on Decision Theory, to be held March 1-3. Speakers will include Steven Brams (New York University, Politics), Dan Farrell (Ohio State), Richard Jeffrey (Princeton), Joseph Kadane (Carnegie-Mellon, Statistics), Brian Skyrms (Irvine) and Paul Weirich (Rochester). All interested parties are invited to attend. More information can be obtained from George F. Schumm, Department of Philosophy, 350 University Hall, The Ohio State University, Columbus, OH 43210.

The Philosophy Department of The Ohio State University is pleased to announce the appointment of James Hughes (Illinois/Chicago) as Instructor for the coming year. The Department is also pleased to announce the appointment of Marshall Swain as Chair, effective September, 1984.

The Philosophy Department of Rice University is pleased to announce the appointments of Assistant Professor Daniel Graham (Texas/Austin); Visiting Assistant Professors John F. Halpin (Illinois/Chicago) and Danny Shapiro (Minnesota); and Mellon Post-doctoral Fellow William E. Schmitz (Michigan).

The Philosophy Department of The University of Texas at Austin welcomes three new faculty members: as professors, Sarah Waterlow Broadie and Hans Kamp will also be associated with the U.T. Center for Cognitive Science, and Lynne McFall as assistant professor.

The Department of Philosophy of the University of Washington (Seattle) is pleased to announce the following appointments: Scott Soames as Associate Professor of Philosophy, effective Autumn Quarter, 1985: Ruth Saunders as Acting Assistant Professor of Philosophy for the academic year 1984/85.

crítica

Vol. XV / No. 44 / México, agosto 1983

SUMARIO

RAÚL ORAYEN, Deducibility Implies relevance? A Negative Answer (II)
ROBERTO J. VERNENGO, Funciones normativas y voluntad de signo
MORRIS LAZEROWITZ & ALICE AMBROSE, A priori Truths and Empirical Confirmation.

Discusiones

M.D. FARRELL, Sayward, el anarquismo y el estado mínimo
JORGE FLEMATTI, The Ontology of Evolutionary Theory

Notas bibliográficas

E. DE OLASO, *Escepticismo e ilustración. La crisis pirrónica de Hume y Rousseau.* [M. Beuchot]
R. BAMBROUGH, *Moral Scepticism and Moral Knowledge.* [J. Esquivel]
C. McGINN, *The Character of Mind.* [A. Rodríguez]
E. MALLY, *Logische Schriften.* [R. Vernengo]

Noticias

Libros recibidos

CRÍTICA, Revista Hispanoamericana de Filosofía is published in April, August and December. Subscription rates are $ 12.00 per year. All correspondence should be addressed to CRITICA, Apartado 70-447, Coyoacan, 04510-México, D.F. México.

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXII, NO. 3, MARCH 1985

MORAL CONFLICT AND MORAL REALISM*

IT has been argued by Bernard Williams,[1] or so I understand him at any rate, that the existence of moral conflicts creates a problem for moral realism. If conflicting moral views are true then the truth of these cannot be absolute.

William's argument has been questioned by, among others, Ruth Barcan Marcus,[2] Samuel Guttenplan,[3] and Philippa Foot.[4] They have argued that even if there are moral conflicts such as moral dilemmas this does not mean that there are moral inconsistencies. I find their arguments convincing. But they all avoid the most serious form of moral conflict. I will present this form in the sequel. I argue that it does not present any problem for moral realism either. I will not argue that moral realism is true, however.

The paradigmatic case of a moral conflict is a moral dilemma. You make a promise but find that, if you keep it, the consequences will be bad. What are you to do? Keep it since, after all, you have promised, or break it, in order to avoid bad consequences? Such dilemmas seldom seem genuine, however. Perhaps the norms in conflict are only first-level rules of thumb; the truth is found at another level of our moral thinking (Hare). Or perhaps the norms only state *prima facie* obligations (Ross). Or perhaps one obligation is lexically superior to the other (Rawls). Or, even though you cannot *decide* which rule is to take precedence, this does not imply

---

*I would like to thank Lars Bergström for helpful comments on an earlier version of this paper.

[1] "Ethical Consistency," *Proceedings of the Aristotelian Society*, suppl. vol. XXXIX (1965): 103–124, and "Consistency and Realism", *ibid.* suppl. vol. XL (1966): 1–22.

[2] "Moral Dilemmas and Consistency," this JOURNAL, LXXVII, 3 (March 1980): 121–136.

[3] "Moral Realism and Moral Dilemma," *Proceedings of the Aristotelian Society* (1979/80): 61–80.

[4] "Moral Realism and Moral Dilemma," this JOURNAL, LXXX, 7 (July 1983): 379–398.

0022-362X/85/8203/0113$00.50

that there *is* no truth about the matter; this move seems to be open to moral realists (Foot). Or your reluctance to give up one norm, in spite of the fact that you realize that the other should take precedence, may have to do with the fact that there is still *some* evidence for it (Guttenplan). Yet, *one* principle and one moral mistake suffices to generate moral conflicts. Even a teleological principle and one future moral mistake can generate *genuine* conflicts, i.e. conflicts that cannot be solved by any of the devices constructed by Hare, Ross, Rawls, Foot, or Guttenplan.

Let us suppose that act-utilitarianism is true and that it yields the following norms:

(1) I ought not to invite a certain friend for dinner.
(2) I ought to invite this friend for dinner and settle my dispute with him.

Both (1) and (2) may well be true if act-utilitarianism is true. Let us suppose that they are. It is true that I ought not to invite my friend because, if I do, then, as a matter of fact, I will not settle the dispute with him but will quarrel with him endlessly. But I ought to invite him and settle my dispute with him. I *can* do so (we assume), and if I don't we will both suffer from the unresolved dispute.

The truth of (1) may be hard to accept in the light of (2). Are we really to adapt our actions to our future moral mistakes? I think we are. If we do so we are obviously faced by genuine moral conflicts, however. In order to avoid inconsistency, therefore, several authors have tried to avoid these conflicts by restricting the application of their moral principles to "relevant" sets of alternatives only[5] or to "maximal sequences of acts"[6] only. Some such restrictions may be acceptable if our problem is to find a rational *method of deliberation* but not if our interest lies in finding out which are the *criteria*

---

[5] This way of avoiding inconsistency by avoiding conflict is advocated by Lars Bergström in "On the Formulation and Application of Utilitarianism," *Noûs*, x, 2 (May 1976): 121-144. Bergström's view is not only inconsistent with moral realism in that it allows for the existence of normative value gaps, it also implies that it is a *value* problem to determine what set of alternatives facing a person in a decision situation is the "relevant" set.

[6] The most recent example of this is Holly S. Goldman. In "Doing the Best One Can," in Alvin I. Goldman and Jaegwon Kim, eds., *Values and Morals* (Boston: Reidel, 1978), pp. 185-214, she restricts the application of the utilitarian maxim to "maximal sequences of acts" and she gives references to others who have done the same before her. These examples are taken from a vivid discussion provoked by Lars Bergström with his seminal work, *The Alternatives and Consequences of Actions* (Stockholm: Almqvist & Wiksell, 1966).

of rightness.[7] It is perhaps rational to reason as if there were normative value gaps; i.e., we may have reason to believe that the effects will be the very best in the long run if we reason that way, but this does not mean that we have good reason to believe that there *are* normative value gaps. The belief that there are seems to be inconsistent with moral realism anyhow. *Each* concrete action is either right or wrong, if moral realism is true. Moreover, there seem to be good *moral* reasons why we ought to adapt our actions to our own future moral mistakes. At least according to consequentialism, what matters morally is what will *actually* happen in the future. J. H. Sobel makes this point in the following way, when he quotes and improves upon Henry Sidgwick in the following manner: "A man's future good on the whole [and his 'subjectively best' action] is what he would now desire and seek on the whole if all the [actual!] consequences of all different lines of conduct open to him were accurately foreseen and adequately realised in imagination at the present point of time."[8] This means that we should adapt our actions to future mistakes made by *others.* We ought not to lend our friend our car if we suspect that he *will* drive heavily drunk, even if he *could* make a good use of it. But, by the same token, we must take our own future mistakes into account. *Both* kinds of mistakes determine what the consequences of our actions will be. Therefore, rather than try to avoid moral conflicts we ought to try to show that their existence is consistent with moral realism.

In my argument for the consistency of moral conflicts and moral realism I take realism about the *future* for granted. This means that I take it for granted that it is true timelessly either that I will or that I will not settle my dispute with my friend. This is *true*, I assume, irrespective of whether we can, even in principle, *find out* (now) whether I will settle the dispute or not. When I say that we ought to adapt our actions to our own future moral mistakes I imply that an action of ours is wrong if, instead of performing it, we could have acted otherwise so that the consequences in the circumstances would have been better; my own future moral mistakes as well as the moral mistakes made by others are then taken to be parts of the circumstances. I don't say that we ought to consider our own future mistakes in our deliberation, nor do I say that we ought not to do so. I hold no view in particular about moral deliberation.

---

[7] Cf. R. Eugene Bales, "Act Utilitarianism: Account of Right-making Characteristics or Decision-making Procedure?," *American Philosophical Quarterly*, VIII, 3 (July 1971): 257–265, for a clarification of this distinction.

[8] Jordan Howard Sobel, "Utilitarian Principles for Imperfect Agents," *Theoria*, XLVIII, 3 (1982): 113–126, p. 123.

If norms prescribing incompatible courses of action like (1) and (2) may both be true, does not this mean that their truth must be somehow relative? Must it not mean that they are true in relation to a certain time or point of view? I think not. For (1) is not inconsistent with (2). Incompatibility does not imply inconsistency. Both (1) and (2) may be absolutely true.

The reason that (1) and (2) are not inconsistent may be thought to be that "the agglomeration principle"[9] is not valid. It does not follow from the fact that I ought to do *a* and *b* that I ought to do *a* (nor does it follow form the fact that I ought to do *a* and that I ought to do *b* that I ought to do *a* and *b*). Such a reason seems somewhat superficial, however. *Why* does not the principle hold? Does not the fact that it is invalid in itself tell against moral realism? We must look deeper into the matter. How do we interpret statements like (1) and (2)?

If moral realism is true, then, most naturally, normative properties (values) like rightness, wrongness, and obligatoriness are *de re*, they are possessed by *concrete* actions.[10] This means that (2) says something like the following. Had I invited my friend and settled my dispute with him, the concrete action, which I had then performed, would have been obligatory (we have taken it for granted that I did not invite him). So (2) is not inconsistent with (1), which says of the concrete action that I perform, that it is obligatory (there was no way I could have acted otherwise during the time interval when I performed it which would actually have resulted, in the circumstances, in better consequences). (1) and (2) are about different actions. This does not resolve the *conflict*, of course. For not only is it true, of the concrete action I perform in not inviting my friend and then doing something else (let's suppose I go to the cinema), that *it* is wrong; it is wrong despite the fact that a (spatiotemporal) part of it is obligatory, i.e., the part consisting of my not inviting him. Isn't that strange? So it may seem. If (1) and (2) were read *de dicto*, if rightness, wrongness, and obligatoriness were conceived of as operations on statements or properties of states of affairs, the reason the agglomeration principle is not true would be rather mysterious. Suppose that my inviting my friend and settling my dispute with him is conceived of as a conjunction of two statements or possible states of affairs, to wit, that I invite him (p), and that I settle my dispute with him (q), and suppose the conjunction of p and q is thought to be obligatory. It would then seem reasonable to argue that, since p and q is obligatory, so are both conjuncts; this seems to follow from a more general principle to the effect that, if p is obligatory and if q follows from p, then q is obligatory as well. If norms are read *de re,* however, the invalidity of the agglomeration principle is self-evident. One cannot agglomerate concrete things in the same way as one can agglomerate statements. No inconsistency is involved in the description of the situation. What is true of a concrete whole need not be true of each part of it. It is true of the chair I'm now sitting on that it has four legs but it is true of none of these legs that *it* has four legs.

---

[9] Cf. Williams, "Ethical Consistency," *op. cit.*, p. 118.

[10] I take it that actions are concrete and particular events produced by wants, desires, or something like that, on the part of an agent. This view, inspired by the writings by Donald Davidson, I defend in "The Morality of Abstract Entities," *Theoria*, XLVI, 2 (1978): 1-18 and in "Responsibility and the Explanatory View of Consequences," *Philosophical Studies*, XLII, 2 (September 1982): 151-161.

Clearly, this argument to the effect that there are genuine normative conflicts and to the effect that this is consistent with moral realism can easily be generalized to cover other moral theories besides act-utilitarianism. W. D. Ross could have come up with the following kind of conflict. He realizes that if he invites a friend for dinner he fulfills a promise he has made; if he does, however, he will, as a matter of fact, insult his guest. He ought to invite his friend and entertain him (not insult him). If he does, he fulfills a promise and violates no rights. He ought also not to invite his friend, for it is worse to insult him than to break a promise (let's suppose). Since Ross himself seems to have held that normative properties (*absolute* rightness, wrongness, and obligatoriness) are *de re* he would have had no problem in combining his moral realism with a belief in the existence of moral conflicts.

To conclude. There are genuine moral conflicts. By doing what we ought to do we sometimes do what we ought not to do. Such conflicts may keep us awake at night, and they may constitute a problem for decision theory. They constitute no threat to moral realism, however.

TORBJÖRN TÄNNSJÖ

University of Stockholm

SUPEREROGATION AND OBLIGATION*

EACH of the following three propositions seems to be true; yet the three together may seem to be inconsistent:

For some effort x, some duty D, and some beneficent act H,

(1) The personal preference not to make effort x, may take precedence over doing act H. (That is, doing H at effort x becomes supererogatory.)

(2) Doing H at effort x may take precedence over doing duty D.

(3) The personal preference not to make effort x may not take precedence over doing duty D.

There seems to be an intransitivity here. That is, where '$\geqslant$' means "may take precedence over," S is some supererogatory act, D some duty, and P some personal interest, $P \geqslant S$, $S \geqslant D$, and yet $\sim(P \geqslant D)$.

This intransitivity involves the claim that we may do a supererogatory act and as a *foreseen consequence* fail to do our duty, and the claim that we may not refuse to make some sacrifice with the *foreseen consequence* that we fail to do our duty. There is another, closely related, intransitivity which involves the claims that we may deliberately violate our duty *as a means* to performing a supererogatory act, but may not deliberately violate our duty *as a means* to achieving our personal goals. The second intransitivity is demonstrated by the following variants on Propositions 1, 2, and 3.

For some personal goal G, some beneficent act H, and some duty D,

(1′) Since the pursuit of G may take precedence over performing H, doing H rather than pursuing goal G becomes supererogatory (S). That is, $G \geqslant S$.

(2′) We may deliberately violate duty D in order to do supererogatory act S. That is, $S \geqslant D$.

(3′) We may not violate our duty in order to achieve personal goal G. That is, $\sim(G \geqslant D)$.

---

[*] This paper is an expansion of a section of a paper written in fall of 1980, first called "Duties and Teleology," and subsequently revised as "The Good, the Bad, and the Obligatory." The problem discussed in this paper arose for me in the course of considering measures of relative strength of positive and negative duties, such as aiding and not killing. The paper was presented in Ronald Dworkin's Political Philosophy Seminar at NYU Law School in April 1983, and I am grateful for the comments of the members of that group. I am most grateful to Thomas Nagel, Mary Mothersill, Derek Parfit, Isaac Levi, and Susan Wolf for their discussion and criticism of earlier drafts. This article is one in a series I have written on aiding; the others published are "The Problem of Abortion," in R. Abelson and M. Friquenon, eds., *Ethics for Modern Life* (New York: St. Martins, 2d ed., 1982); "Equal Treatment and Equal Chances," forthcoming in *Philosophy and Public Affairs* (Spring 1985), and "Killing and Letting Die: Methodology and Substance," *Pacific Philosophical Quarterly*, LXIV, 4 (October 1983): 297–312.

0022-362X/85/8203/0117$02.10

In this paper I will try to show that, even within a basically nonconsequentialist ethical system, all six propositions can be true. I will also try to show how it is that "may take precedence over," as well as certain other relations, are not transitive. In particular, although it is ordinarily thought that we must do what, all things considered, we have a duty to do, I will argue that it is sometimes permissible to do a supererogatory rather than a dutiful act.

I will also argue that there are at least two different measures for the importance and stringency of acts. The first is the amount of effort we have to make to do an act, the second is whether doing one act may override a requirement that we do another act. The fact that there are these two measures of stringency and importance raises the question of whether we may, in fact, refuse to make great efforts to do an act (e.g., our duty *D*) which may be superseded by another act (e.g., supererogatory act *S*) for the sake of which we need not make such efforts. I shall argue that we do have to make certain efforts to do our duty, though we need not make the same efforts to do a supererogatory act which we may nevertheless do instead of the duty. Furthermore, we may not violate a duty for some personal goal, though we may pursue the same personal goal instead of doing a supererogatory act for the sake of which we may violate the duty.

I

It is commonly thought that we sometimes have a duty to help people, at least when the sacrifice to ourselves is not very great. It is also commonly thought that if the efforts required of us to help are high, helping is supererogatory, i.e., beyond the call of duty. The personal preference not to make the large effort, or to pay the large cost, may outweigh considerations in favor of helping. Therefore, proposition 1—that personal preference not to make effort *x* may outrank the beneficent act *H* and make it supererogatory—is commonly accepted as true. (Likewise, if I must sacrifice an important personal goal to help someone, helping is commonly considered supererogatory. Therefore, proposition 1′ is also commonly accepted as true.)

Now consider the following case: I have promised to meet someone at 12 P.M. for lunch; so I have a duty to be there. I shall use the notion of duty so as to allow that there might be other duties of mine that conflict with my duty to go to lunch. In cases of conflict I would have to decide which was my strongest duty, and then *it* would be my duty all things considered. (Alternatively I might refer to my conflicting duties as "prima facie duties," and to the strongest one as "my duty.") Now imagine that on my way to the lunch appointment I come across a car crash. One of the victims has had his kidneys crushed, and he needs an immediate kidney transplant in order to survive. I am the only person available who has the right type of kidney. Given the size of the sacrifice to me, most would say that it is not my duty to donate one of my kidneys. Therefore, in these circumstances (barring the existence of any other duties) the only *duty* I have is to keep my lunch appointment.

Suppose, however, that I am willing to give up a kidney, but this will cause me to miss my lunch appointment. It still seems that my only duty is to keep the lunch appointment. Nevertheless, if I want to give the kidney I may, and I need not do my duty. Indeed, it would be morally stupid of me to think that I had to do my duty. I would be a moral idiot (though not immoral) if I felt that I had to keep my lunch appointment.

Those who think that I have a duty to save lives when the cost to me is small, and also a duty to keep my lunch appointment, would, no doubt, think that the duty to save lives outranks the duty to keep lunch appointments. If they also maintain that I must always do my strongest duty before doing supererogatory acts, they will be committed to the view that, if there is a choice, I am obliged to do what will save someone's life at minimal cost to myself (a duty), rather than make great sacrifices that I am willing to make in order to save a great number of lives. Thus, it may be my duty, for example, to give a penny to save the life of one person on an island, rather than give away a fortune I am willing to give away to save thousands of lives on another island, when I can reach only one of the islands in time to save any lives. This seems quite the wrong conclusion to reach.

I conclude that, if "I am obliged to do act *y*" implies that doing others acts would be wrong, then I am not always obliged to do my strongest duty. If 'I am obliged to do *y*' just means "*y* is my strongest duty," then I need not always do what I am obliged to do. Contrary to the views of many, there is no strict lexical ordering between duties and acts of supererogation, such that we may perform a supererogatory act only if does not involve our failing to do our duty.[1] We cannot always claim that we are prohibited from doing something that would require much sacrifice on our part and produce a great good, because we are bound by our duties. Therefore, though supererogatory acts are beyond the call of duty, it should not be thought that doing them implies that we have first done our duty and then gone beyond it. In doing a supererogatory act we may have permissibly failed to do our duty. As proposition 2 states, an act of supererogation may take precedence over a duty. (Likewise, if I wish to save a life, I may deliberately violate my duty not to take someone's car without his permission, as a means to save the life, even though, because of the risk it will involve to me, this is not a life it is my duty to save. Therefore proposition 2′ is correct.)

---

[1] The view that there is a lexical ordering is held by R. M. Hare [*Moral Thinking* (New York: Oxford, 1981)], David Heyd [*Supererogation* (New York: Cambridge, 1982)], and David A. J. Richards [*A Theory of Reasons for Action* (New York: Oxford, 1971)].

Is there a way of avoiding this conclusion that duty does not take automatic precedence over supererogation? Objections might be made to the argument I have presented. The first objection might be that someone who is *willing* to make a large sacrifice is therefore obliged to make it, and so does have a stronger duty than his luncheon appointment. But I do not think it is true that being willing to make a large sacrifice implies that one has an obligation to make it.[2]

A second possible objection is that since my friend would himself be obliged to release me from my duty to him, once he found out how much good I could do if I missed lunch, I do not have a duty to meet him that is overridden by a supererogatory act. I believe this objection is also wrong, whether or not my friend would have a duty to release me. It is wrong because we can imagine cases in which it would be *supererogatory* for the person to whom I am duty-bound to release me, and yet it would be permissible for me to fail him in order to do a supererogatory act. For example, suppose that I have a choice between (a) giving up my own life to save a thousand people (which is not my duty), and (b) saving one person at small cost to myself (which is my duty). The loss that the one person will suffer if I save the thousand instead of him is larger than he would *have* to make in order to save the thousand people. Therefore, in this case, the one person would not have to release me from my duty to him. Yet I think that I may decide in favor of the thousand though it is supererogatory for me to do so. (I believe I may save the thousand even if I have promised to save the one.)

A third possible objection is that duties have escape clauses built into them, so you have a duty to keep your luncheon date except when. . . . This view of duties as fully specified (like the comparable view about rights[3]) is problematic, since it implies that, given that we never plot out all the exceptions in advance, we never know what our duties are. Also, it has difficulty accounting for compensation that we may owe for having failed to keep our appointment, since it implies that we have not failed in any duty. Further, the doctrine of fully specified duties really seems to have all its work done for it by a prior weighing of a simple duty (e.g., to keep the lunch appointment) against other acts (e.g., supererogatory acts), and so is parasitic on the doctrine that a duty may be overridden by a supererogatory act. Full specification makes the most sense (and does not confront the issue of compensation) when we are dealing with a duty to aid which is uncontracted and not dependent on a special status: each person may have a duty to save lives at moderate personal cost except when he wishes to do more good even at large personal cost.

---

[2] I have discussed this point in greater detail in a longer version of this paper, which I hope to use elsewhere.

[3] As discussed by, for example, Joel Feinberg in "Voluntary Euthanasia and the Inalienable Right to Life," *Philosophy and Public Affairs*, VII, 2 (Winter 1978): 93-123, and Judith J. Thomson in "Self-Defense and Rights," The Lindley Lecture (Lawrence: University of Kansas, 1976). Thomas Scanlon suggests specified duties.

Finally, it is possible to argue that supererogatory acts do not override duties since our *general* duty is to do *either* one of our *particular* duties (e.g. keep the lunch appointment) *or* do a supererogatory act.[4] It may be objected to this view that there are other reasons for not doing one's particular duties, that it, like fully specified duties, makes it difficult to account for a duty to compensate when one fails to do a particular duty, and that it is too consequentialist, implying that one may always abandon one's duties in order to do supererogatory acts. It is also problematic since it seems to imply that if I fail to do my particular duty in order to do a supererogatory act, I am then obliged to do that act. But although the pressure on me may increase to do the supererogatory act, I do not believe that failing my lunch date will necessarily leave me with a duty to give up my kidney to save someone. It would, however, be sufficient for my purposes just to argue that a supererogatory act may supersede a particular duty.

The point is that if it would have been our duty in the circumstances to save a thousand people if the cost to ourselves had been small, even when doing this would have meant that we could not fulfill the duty to save one person, then we may save the thousand people even when, because it involves a much greater cost to us, it is supererogatory to do so. Furthermore, the fact that we may do the supererogatory act does not mean that those whom we abandon are obliged to make the large sacrifice of releasing us from our duty to them.

The supererogatory acts I have discussed are of one sort: they involve acts that would have been our duty, given a general duty of beneficence, if they did not require large efforts on our part. The justification for not doing our duty seems to derive from the ability of this duty of beneficence to dominate another duty. Its ability to outrank the other duty becomes salient when the large sacrifices required to carry out the supererogatory act are voluntarily undertaken. But it is important to note that acts whose supererogatory nature does not depend on their requiring large efforts—e.g., acts of kindness, mercy, forgiveness—might also be done instead of duties, even though the person to whom we have a duty is not obliged to release us from it.

---

[4] This point was suggested by Isaac Levi and Susan Wolf.

A case that would satisfy these conditions is one in which: (a) there is a noncrucial large gain to the recipient of the supererogatory act, (b) the act costs me nothing to do, and (c) the act leads me to fail in my duty to someone else, who thereby suffers a loss that is not crucial but is large enough so that he would not have to release me from my duty to him once he saw how I would otherwise act. Suppose, for example, that someone asks me to help him complete a contest application which he must submit this evening. I am very good at contests, and the low-income man stands to win a lifetime secure income if I help him. I have, however, already promised to spend the evening with a sick friend who really does need the company. It seems to me that I may leave my friend to help the man earn his lifetime income though it is a supererogatory act.

II

I have argued that we may sometimes fail to do our duty in order to do a supererogatory act (i.e., that proposition 2 is true). I believe that the truth of proposition 2 implies that it is sometimes the case that we may fail to do an act (our duty), which we might ordinarily be obliged to make *large sacrifices* to do, in order to do another act (the beneficent act) which we wouldn't be obliged to make those same large sacrifices to do. For example, if I am a contractor and have promised to give my construction plans to a business associate by a certain date, I may be obliged to spend a great deal of extra money (e.g., on special delivery, keeping my employees working overtime, etc.) in order to meet the promised deadline. (Call this the *contractor* case.) I am not obliged to make the same effort in order to save someone's life, when I have not promised to. Yet it would be morally permissible for me to miss the business deadline (e.g., to skip the meeting at which I was to give the plans) in order to save someone's life.

In general, we seem to feel obliged to go to a great deal more trouble, for example, to keep our professional commitments than to aid people in need. Yet we feel that we may drop these professional commitments to save a life. Proposition 2, therefore, is true not only when the duty we fail to do is one that we would *not* have had to make great efforts to do. Indeed, when the effort required to save a life is small, we may have a *duty* to save the life, rather than meet the business deadline, although, in general, we would have to make *greater* efforts to keep the business deadline than to save the life. Furthermore, we may skip the business meeting to save the life, even if we wouldn't be willing to make large efforts to save the life should this be necessary and if we would be willing to make the large efforts to keep the business meeting.

This contractor case shows that there are two possible standards for the stringency or importance of an act. One is the *efforts standard.* Keeping our business commitment seems to be a more stringent duty than savings lives *in general,* in the sense that we are required to make greater efforts to fulfill it than we are required to make to save a life. That is, some efforts that would be beyond the call of duty in order to save a life would not be beyond the call of duty to keep a business commitment. The efforts we have to make to keep commitments can be seen as *subsidiary duties.* Someone who promised to meet me, but went to no trouble at all to keep the promise would be failing in his duty. Some duties may *require* a great deal of us. For example, if I have a duty not to kill someone, I may have the subsidiary duty to make a great effort to avoid killing him. (How much I am *willing* to do, both of what I am required to do and of what I am *not* required to do, in order to fulfill my duty is a mark of how conscientious I am, how much I *care* about my duty.)

The second possible standard for the stringency or importance of two acts determines which act may or must be done when it is impossible to do both. (Call this the *precedence standard.*)

Proposition 2 implies that sometimes we may do (or even must do) a type of act that is required of us less stringently when measured by the standard of efforts required to perform it, rather than do a type of act that is required of us more stringently by the efforts standard. Put more formally, where B = promise keeping, and A = saving a life, $>$ = dominates, and $-$ = not doing, the following might be true:

(I) We have to make effort x for the sake of doing B (i.e., $B > -x$; doing B dominates avoiding the effort).

(II) We are not required to make effort x for the sake of A (i.e., $\sim(A > -x)$; it is permissible either to avoid the effort or to make the effort to do A). And yet,

(III) (a) We may do A rather than B [$\sim(B > A)$; i.e., we have the option of doing A or B], or

(b) We must do A rather than B (i.e., $A > B$).

On the measure of efforts required, B seems to be more stringently required than A; yet, on the measure of which we must do if given a choice, either they seem equal or in some cases (when we must do A), A seems more stringently required. We have here generated the intransitivity described at the beginning of this paper. Given a duty D, personal interest (in not doing x) P, supererogatory act S, and where '$\geqslant$' means "may take precedence over", (I) states that $\sim(P \geqslant D)$, (II) states that $P \geqslant S$, and (III) implies that $S \geqslant D$; or equivalently, $P \geqslant S$, $S \geqslant D$, yet $\sim(P \geqslant D)$.[5]

One way out of the intransitivity is to deny that we must make great efforts to do our duty. The fact that we may do the act required less stringently as measured by the efforts-required standard, rather than the more stringently required type of act, may even lead one to reason as follows: If one may (and sometimes must) break one's business promise in order to save a life [$\sim(B > A)$] and if one may refuse to save a life in order to avoid large efforts $x[\sim(A > -x)]$, then, perhaps, one *may* break one's business promise in order to avoid large effort $x[\sim(B > -x)$ or $P \geqslant D]$. The idea here is that if I may or must do A rather than B, they must be essentially equal in significance or else A must be even more significant; so, if I needn't make effort x (sacrifice myself personally) to do A, I needn't make the effort to do B. Another way to put the point is that if we have a choice of doing our duty or doing the supererogatory act (proposition 2), and a choice of doing the supererogatory or conserving our effort (proposition 1), why don't we have a choice between doing our duty and conserving our effort (denial of proposition 3)?[6]

If we reject this suggested solution, as I think we should, we must explain the fact that the act that is less stringently required by the efforts measure may take precedence, and we must also show that, even if we are not required to make great efforts to do it, we may still be required to make just such efforts to do the duty over which it may take precedence. Before attempting this, however, it will be useful also to make clear how the second version of the intransitivity described at the beginning of the paper arises.

---

[5] Because an act that is less stringently required by one measure may take precedence over another act that is more stringently required by that same measure, proving that negative duties are more stringently required than positive duties in the sense that we must make a greater effort to perform them, does not resolve the problem of whether we may violate a negative duty to perform a positive one. For example, Philippa Foot says that we may not kill one person to save others, because the duty not to kill is more stringent than the duty to aid [in "The Problem of Abortion and the Doctrine of Double Effect," reprinted in Bonnie Steinbock, ed., *Killing and Letting Die* (Englewood Cliffs, N.J.: Prentice-Hall, 1980)]. But that greater stringency may mean only that we would *have* to make a greater effort to avoid killing someone than to aid others, and this does not imply that if we *want* to we may not violate a negative duty in order to aid .

[6] Another way out of the intransitivity, of course, is to deny that aiding others at large cost to ourselves is supererogatory. An in-between position is to argue that making a large sacrifice in order to aid someone is at least not supererogatory when we have already failed to make an obligatory large sacrifice to perform our duty. I believe both these positions to be indefensible, but omit further discussion of them for reasons of space.

It has been argued that we may sometimes violate a duty as a means of doing a supererogatory act (proposition 2′, $S \geqslant D$). For example, we may take someone's car without his permission to save a life at great cost to ourself. It has also been said that it is commonly thought that we need not sacrifice goals that are personally important to us in order to save life (proposition 1′, $G \geqslant S$). But, if we may violate duties to save lives, why may we not violate duties for the sake of what may take precedence over saving lives, our personal goals? Why may we not take someone's car without his permission to go on our vacation? [That is, why isn't proposition 3′, $\sim(G \geqslant D)$ false?[7]] How can we show that we may have leeway to violate duties for the sake of an act for which we needn't sacrifice personal goals, but do not have such leeway to promote those personal goals?

I shall begin by dealing with the first intransitivity, since the answer to it will also help solve the second. My strategy will be to consider possible explanations of the permissibility of doing the supererogatory instead of one's duty, and then see whether the correct explanation provides a basis for distinguishing between doing the supererogatory act and avoiding personal sacrifice as a ground for failing to do one's duty.

III

The first suggested explanation for the permissibility of doing the act less stringently required by the efforts standard is that saving-someone's-life-at-small-cost-to-oneself is more stringently required by the precedence standard than keeping-the-business-appointment-at-great-cost-to-oneself (with the emphasis on the different costs involved for oneself). But this will not do as an explanation; for what if there is a great cost attached to saving someone's life but a person is willing to pay that cost? The willingness to make large sacrifices to save a life would not make it impermissible to save the life rather than keep the appointment. Second, there may be no great cost involved in keeping one's business appointment on some occasion, and one could nevertheless choose to save the life, even at great personal expense. Third, this explanation does not tell us why we *may* do a less stringently required type of act at great expense; it will only tell us why we *must* do it at small expense.

---

[7] Such a question comes to mind as a possible criticism of the position for which Samuel Scheffler argues in his *The Rejection of Consequentialism* (New York: Oxford, 1982). Scheffler claims that we have the prerogative not to make great personal sacrifices to do what is best over-all from the objective point of view. He also suggests that it may be permissible to kill one person in order to save the lives of five others. But if we may kill for the greatest good, and self-interest may take precedence over the greatest good, perhaps we may kill a person to promote our self-interest. Would Scheffler permit this? Notice that even if the assumption of transitivity, on which this argument is based, is wrong, Scheffler's views may still lead to the same conclusion. This is because he rejects a moral distinction between harming and not aiding. The absence of such a distinction, combined with the permissibility of not suffering large losses to aid, would imply the permissibility of killing in order to avoid large losses. For example, if I may carry on my work rather than save a life, I might run over someone as a necessary side effect of doing my work. I have discussed Scheffler's views in "Prerogatives and Restrictions," presented at the meeting of the APA Pacific Division, March 1985, and the permissibility of my killing one to save many (both when the many will die because *I* set up a threat to them and when someone else threatens them) in "Constraints and You," presented at the APA Pacific meeting, March 1984.

Fourth, to say that we may do what is less stringently required by the efforts standard just because it now involves less effort to ourselves suggests that acts that are in general less stringently required may be done when it is in our self-interest to do them, given the alternative. But this would imply that we may devote time to volunteer work whenever we can thereby escape some strenuous commitment, and this is wrong.

Furthermore, doing the act less stringently required by the efforts standard, instead of one's duty, may not release one from having to make great efforts to compensate the person to whom one had the duty; yet doing the less stringently required type of act does not become impermissible just because efforts to oneself increase when it is possible to compensate.

A second explanation of the possibility of choosing the type of act less stringently required by the efforts standard is that although, in general, the more stringent type of act produces more good than the less stringent type, this is not always so. On some occasions the more stringent type produces very little good, and the less stringent type produces much more good. In these situations we may fail to do the more stringent type of act.

This sort of explanation will not do for cases in which the more stringent type of act *does* produce considerable good and is such that we would be obliged to make large efforts to do it if only we did not have the opportunity to produce more good by doing the supererogatory act. In particular, it doesn't seem that the proposed explanation can account for cases in which we could omit a duty that would accomplish something nontrivial, something the duty was ordinarily intended to produce. Further, it does not *seem* to be true that, in general, acts more stringently required by the efforts test do more good than acts not so required, or that if most people did the latter instead of the former things would be much worse.

A third possible explanation of why the type of act less stringently required by the efforts standard may be done is that the more stringently required type of act is a *minimum*, but minima are *not necessary* requirements. It is right to require people to make great efforts rather than fall below a minimum. However, they may do something else that is better than the minimum, as a substitute for it, even though they wouldn't have to make great efforts to do the substitute and even though the substitute does *not encompass* the minimum. By requiring great efforts to do it, we emphasize that a type of act is a rock bottom below which we may not fall. Other types of acts are beyond the minimum, and so we do not demand as much of people to do them even though they produce better overall consequences. But this does not mean that doing such an act is not an adequate substitute for doing the minimum, in some cases. To use an example from another area: We are required to work hard in school to learn to read and write, not so hard to compose great poetry. But composing great poetry (and dictating it to a machine) may nevertheless be more than a fair exchange for learning the minimum. In a sense, the fact that we do another act that produces much good means that we haven't really fallen below the minimum, that it is only if we fail to read and write *and* also fail to do poetry that we are below the minimum.[8]

A fourth explanation of the phenomenon in question is that the difference in how stringently different types of acts are required as measured by the efforts standard is not a function of how important or good what they accomplish is. It is simply that only certain types of acts are of the sort that can be required of people even at large expense to themselves. Other types of act cannot be required of people, or cannot be required of them at great expense, even if these acts have good and important consequences. In other words, being this *requirable* type of act is not necessarily or only a function of having better or more important consequences than other types of acts would have. It is not necessarily preferable that you do one type of act rather than the other; it is just that it is permissible to require you to do only one type of act at great expense to yourself. Similarly, a prisoner may be required only to serve his term; he may not be required to participate in drug research. However, if he does participate, we may prefer it, and release him from his prison term. (I shall not here attempt to describe what makes an act requirable at great expense. But it is important to note that there are at least two possible senses of 'not requirable'. One is that we are not entitled to require something of someone; the second is that it is not feasible to require something of someone, though, if it were, we would be entitled to. I am concerned with the first sense of 'not requirable', but it might also be argued that when we do not require big efforts for acts with very good consequences it is only because it is not feasible to enforce the requirement.[9])

---

[8] A variation of the third explanation would insist that we may allow some people to fail to do the minimum and do the supererogatory act instead, so long as most do the minimum.

It seems that we may not require effort from an agent merely in accord with the goodness of the consequences of doing so. But it is worth noting again that an agent may do a supererogatory act because it produces important consequences though the costs to those he lets down in failing to do his duty will be greater than they would have had to agree to pay. This indicates that there are two different types of costs extractable for doing acts: dutiful acts may justifiably impose costs on the agent who does them; some supererogatory acts may justify the agent in imposing costs on someone else. What might account for this is a view that sacrificing oneself is more difficult than sacrificing others. It may be, however, that the only large losses an agent may impose on others against their will are the consequences of his not doing something for them (even something that they had a right that he do). It may be wrong to take from them something significant which is theirs by right and which they can have without his help.[10]

This fourth explanation really includes the third as a part, since the third explanation identifies "what we may require people to do" (even at large cost to themselves) as "the minimum." Both the third and fourth explanations, however, share the defect of being too consequentialist. That is, they imply that duties are minimal or requirable acts that *never* stand in the way of acting for the greater good. But the nonconsequentialist view is that it is not the case that just any greater good can serve as a reason for failing to do one's duty.

A fifth suggestion[11] as to why we may do a type of act less stringently required by the efforts standard focuses on the costs of *not* doing the supererogatory act. On the efforts measure of stringency, I may have to pay a high cost to do a duty. But this does not necessarily mean that I must pay every cost to do my duty. For example, I may have to take a plane to keep my business appointment, but not give up my life. If I am willing to save lives, though it is not required that I do so, the cost to me of not saving the lives (not doing what I want to do) must be added as part of the cost to me of doing my duty. This may make the cost of doing my duty higher than can be required of me, and I may be released from doing it. If I weren't willing to save the lives, the cost to me of doing my duty would not be so great.

---

[9] Derek Parfit made this suggestion.

[10] I have discussed this distinction in papers on the killing/letting die and harming/not aiding distinctions.

[11] Made to me by Thomas Nagel.

This suggestion seems to remedy one of the defects of the third and fourth suggestions, since it would explain why not every act that produces a great good may override a duty. It is only if the loss to me of the good that would come from the supererogatory act is greater than the cost I would have to bear to do my duty that I may abandon my duty.

The emphasis in the fifth explanation is on the cost *to me* if the many lives I am willing to save are lost. That this is a mistake, the following may help show. Some duties may require a very great deal of us. For example, I may be required to give up even my life to fulfill some duties (e.g., if someone threatens to kill me unless I kill someone else, I may have to die rather than kill someone else). Suppose there were a conflict between my carrying out a duty for which I *could* be required to sacrifice my life, and my saving many lives. The fifth explanation could justify my doing the supererogatory act rather than the duty, when doing the duty *in fact* costs *nothing other* than the loss of the many lives, only if I cared significantly more about the combined loss of the many lives and of my own than about the loss of my own life alone. This is the only way their deaths could raise the cost *to me* of doing my duty beyond the cost I am already required to pay. (This does not mean that I would have to be willing to give up my own life to save the many people in order to be justified in abandoning a duty that I could be required to die for.)

I believe this implication of the fifth explanation is wrong. First, suppose I were required to give up *everything* I most care about, even those things which mean more to me than my life, in order to do some duty. Does this mean that I cannot abandon this duty in order to save many lives? I believe not, since it is possible for someone to be willing to sacrifice anything he cares about in order to do his duty, and yet not be willing to allow other people to die just so that he may do his duty. Indeed, he may care more about doing his duty (as measured by how much he would be willing to sacrifice to do it) than about saving many lives, and yet still abandon his duty to save the lives. He may be willing to give up his life to do his duty and yet fail to do his duty in order supererogatorily to save many people, even though he wouldn't be willing to give up his life just to save those other people—and though he would give up his life to save those people *if* by doing this he could avoid sacrificing his duty in order to save those people.

How does such a person reason? I believe he compares the objective importance of the content of his duty with the objective importance of many lives being lost, and finds the latter to be much greater. Even though he would give up more of what has only personal value to him in order to do his duty than he would give in order to save the many lives, he cannot allow something of much greater objective value (many lives) to be lost just so that something of less objective value (the content of his duty) be achieved. In general, on all measures of how much a person cares for something, he may care more about the consequences of doing *M* than of doing *N*, and still reasonably do *N*. So although someone *may* fail to do his duty in order to do a supererogatory act because he cares more about its consequences (saving many lives), he may also do one supererogatory act rather than another (e.g., saving a hundred strangers rather than ten friends) because it seems the objectively better thing to do, though he doesn't care as much about the outcome.

When someone would sacrifice what he cares about more than his own life to produce a good for whose sake *alone* he would not give up his life, he *will* still sacrifice his life to produce the good if this prevents the loss of what he cares about more than his life. What this shows is a compromise between the objective and personal points of view. He does not sacrifice his own life because he directly compares it in an objective evaluation with the great good. Rather, this sacrifice is mediated by its role in saving what has greater personal value to him, regardless of whether this is objectively more important than he is. This account implies that an agent may refuse to weigh his own life on an objective scale and yet weigh on that scale something he cares about more than his own life. Perhaps this is an indication of how closely being an agent is tied to not seeing oneself as just another creature in the world.

This objection to the fifth explanation shows that when I am willing to save lives supererogatorily, the cost of doing my duty instead should include the objective value of the lives that will be lost (where objective value may include the value of their lives to those who would die if I do not aid). (We may call the fifth explanation so revised "the revised fifth explanation.") It should not merely include the cost *to me* of those deaths. If I were *not* willing to save the many lives supererogatorily, the cost of doing my duty should not include the value of those lives.

Including the objective value of what will be lost, not just the value of the loss to me, in the cost of doing my duty, however, reintroduces the problem of utilitarian calculation. That is, how can we account for the fact that doing much more good may override doing our duty, but doing only *somewhat* more good may not? We could account for this fact if we saw the cost of *not* doing the supererogatory act as the cost-to-me, since the costs I am required to make to do my duty are assumed to be fixed and when the actual costs go higher, I needn't do my duty. But if it is not a question of the costs to me being too high, we shall need an independent measure of what objective losses may and may not be sacrificed in order that a duty be done. I will not settle here the issue of how much good is necessary to override the duty.

Suggestions 3, 4, and 5 all attempt to explain why we *may* do the type of act which is less stringently required by the efforts standard. None of them, however, explains why we sometimes *must* do the act that is less stringently required. For example, none of them explains why we *must* save a life at small cost to ourselves rather than keep an important business appointment.

To explain this, we may have to return to the first proposed explanation and accept a view that combines each act with the effort necessary to do it, and weighs the stringency of this pair against other such pairs. We should do this without implying, however, that we must do whichever act requires less of us *because* it requires less of us. So, saving-a-life-at-small-cost-to-us may be required as much or more than keeping business promises, and the greater good produced by the former makes it dominate the latter.

It seems that we can explain to some degree how an act can be more stringently required as measured by personal efforts demandable to do it, and yet be dropped in favor of doing an act that is less stringently required by this measure. We may be *required* to make large sacrifices for some types of acts (e.g., minima) but not others, even though the latter have more important consequences. However, if we are willing to make sacrifices to do the act that has the much better consequences, then we may abandon the more stringently required act (our duty), when either (a) the cost to us of not doing the supererogatory act is greater than we would have to pay to do our duty, or (b) the cost to the world (as opposed to cost to us) of our not doing the supererogatory act is much greater than the cost that would be required to do the duty. Furthermore, sometimes we *must* abandon an act that is more stringently required as measured by efforts standard in favor of an act less stringently required by the same standard. This is because relative stringency can also be measured by combining acts and the efforts necessary to do them, and it may be more stringently required that we do acts at small cost that we needn't do at large cost, than that we do acts that we must do even at large cost.

This explanation of why we may fail to do our duty in order to do the supererogatory act enables us to concede that (1) personal preference *P* not to make effort *x*, may take precedence over supererogatory act *S* (proposition 1), (2) doing supererogatory act *S* may take precedence over duty *D* (proposition 2), and at the same time accept that (3) personal preference not to make effort *x* may not take precedence over doing duty *D* (proposition 3). The explanation of the intransitivity of "may take precedence over" depends on the distinction between requirable and nonrequirable types of acts and on the fact that we need not suffer every loss of valuable consequences for the sake of doing a more stringently required act.[12]

IV

The second question we had to answer was whether the permissibility of intentionally violating a duty as a means to promote the greater good, and the permissibility of pursuing personal goals instead of the greater good, implied that it might be permissible to intentionally violate a duty to promote one's personal goals? The answer proposed above to the first question also provides possible grounds for answering "no" to this second question.

There is, however, an additional point worth making which relates to both questions but which arises more naturally in considering what means we may use to achieve various goals. It involves focusing on the choice between personal good and a supererogatory act, rather than on the choice between duty and a supererogatory act which was emphasized in the explanation discussed above. As noted above, our right to require people to make large efforts or to forego personal goals in order to do their duty may not depend only on how much good results from their doing their duty. For example, we may be able to require greater sacrifices to keep a business promise than to save a life. Likewise the prerogative (or dispensation) to avoid inconvenience and pursue one's own goals rather than the greatest good, is not based on the claim that our personal interests are objectively more important than the many lives we could save. But the permissibility of abandoning or violating a duty may well depend on the objective importance of what we hope to achieve by doing so.

---

[12] An addition will be made to this explanation later (136).

It seems reasonable to think that the permissibility of supplanting one goal (great good) with another (personal good) is more likely to lead to the permissibility of violating the same type of duty for either goal if the goal of personal good were as important as the goal of greater good for which we may violate the duty. (I will examine this reasonable belief further below.) So long as the prerogative to choose the personal goal is not based on its being as or more important than the goal of greater good, its ability to replace the greater good does not license use of the same means to accomplish it.

One may be misled by thinking that the permissibility of avoiding the personal sacrifice or the permissibility of pursuing personal goals rather than saving many lives is based on the dominance of the personal concerns in virtue of their greater importance. But neither are they more important, nor do they dominate in the sense that we *must* pursue personal interest rather than the greater good. We have a *choice* whether to pursue personal good, not a requirement to pursue it.

In general, even if it is permissible to do x or y and if it is permissible to use a given means to achieve x, this will not mean that we may use the same means to achieve y. For example: I have a choice between going swimming or hiking. If I go swimming, I have a right to have a key to the pool, but this does not mean that I have a right to the key if I go hiking.

In this case, of course, the key to the pool would not help me to achieve my new goal (hiking). Would the general point hold even if the same means would do me good with either goal? Consider another example: I may do personal chores (personal goal) or go and do volunteer work (greater good). My neighbor will understand if I take his car to do the latter, but not to do my personal chores, though the car would be useful for that purpose too. Is my neighbor's position inconsistent? It certainly doesn't seem to be.

I have claimed that the relation between (1) the pursuit of personal goals and avoidance of personal efforts and (2) supererogatory acts, is neither a strict dominance relation nor based on the greater objective importance of the personal. I have also suggested that it seems reasonable to think that if the pursuit of personal goals were equal in importance to the consequences of supererogatory acts, it would be easier to justify violating a duty to pursue the personal. But I believe it is important for understanding the intransitivities discussed in this paper to examine further what would be true *if*, first, the relation between the personal and supererogatory were one of dominance, and, second, if the dominance were based on importance.

Suppose that there were a strict dominance relation between personal concerns and supererogation, and between supererogation and duty, in the sense that one *had* to do the supererogatory act rather than one's duty, and one *had* to put one's self-interest ahead of the supererogatory act. To have such strict dominance is to take acts with the same content as those currently considered supererogatory and, in part, *forbid* doing them at large personal cost. This gives one a duty to care for oneself rather than do acts beyond the call of duty. It also takes acts with the same content as those currently considered supererogatory and *requires* that they be done rather than a duty. This requirement turns the act currently considered supererogatory into a duty in some cases. (To be consistent with the prohibition on sacrificing oneself, the "supererogatory" would have to be such as not to involve a large sacrifice of one's interests, e.g., mercy.)

My claim is that even such a dominance relation would not imply that it was permissible to violate one's duty for the sake of personal interest. This is so even though it seems like a straightforward requirement of transitivity that personal interest dominate duty. This means that "ought to do ___ rather than . . ." and "obliged to do ___ rather than . . ." are intransitive relations. If we let 'S^*' stand for the content of the act currently thought to be supererogatory, and '$>$' for dominates, the intransitivity we will be investigating is how $P > S^*$, $S^* > D$, and yet $D > P$. One explanation for this intransitivity is that there could be different reasons for each of the dominance relations. The following is an analysis, in keeping with the explanation given above for the intransitivity between P, S, and D, of how this could happen. S^* could dominate D because it produces better consequences, P could dominate S^* because of the value of someone devoting himself to self-realization in comparison to the value of some better over-all consequences, and D could dominate P because of the greater importance of accomplishing certain minimum requirements in comparison to the value of devotion to self-realization.

The following is a somewhat different analysis of how strict dominance is compatible with intransitivity when duties involve personal commitments. S^* could dominate D for one reason (better over-all consequences), P could dominate S^* for a second reason (greater value of devotion to self-realization), and it could be that neither these reasons nor others lead to the dominance of P over D because, for example, self-realization may not take precedence over keeping commitments that the self has voluntarily undertaken. In this second analysis what is emphasized is how in the case of duty one aspect of the self (its voluntary commitments) overrides *another aspect of the self* (its interests and potentials). It might be over-all destructive to the self for its commitments to be sacrificed to its interests. The commitment aspect of the self, however, may be sacrificed *to over-all good* without this damaging the internal relations between aspects of the self. Likewise, not doing supererogatory good when there has been no commitment to do it, would not create a conflict between aspects of the self.

The possibility that there are different reasons for each of the dominance relations contrasts with the view that there is a single scale, e.g., production of happiness, according to which we decide about each dominance relation.

Suppose we now describe the dominance relation explicitly in terms of importance: We must do S^* instead of D since it is more important; we must do P rather than S^* because it is more important. Still, it may be that D is more important than P. I believe the following way of explaining this intransitivity (with the help of the following diagram) is especially illuminating for understanding duty. Where '>' means "dominates,"

$$S^* > \underparen{\overparen{D > P}} > S^*$$

The diagram emphasizes the special relationship holding between my duty and my self-interest, how tightly they can be bound together in a dominance relation, while individually duty and personal interest go their separate ways, so to speak, with respect to S^*. It might be that devoting oneself to one's development is morally more important than S^*. On the other hand the content of personal duties might make them unimportant morally in comparison with S^* (e.g., great good of the world). Indeed, D might be sacrificed to S^* on a strict utilitarian calculation. Yet the unimportant personal duty could make claims on me that override my self-development. In virtue of its being my duty, something whose content is in itself not very important (D) could "humble" something that is in itself very important (P). What is noteworthy is that D never be subordinated *to P*, whatever else it is subordinated to, and whatever else *P* dominates.

An analogy may help to understand this point. It might be part of a theory of royalty that a king must never sacrifice his personal interests for the greatest good of his country; he must be its single venerated subject. He may, however, have to fail in his duties to his parent and even sacrifice the parent for the greater good of his country. (He may not care about his parent much, so her fate is not among *his* interests.) Yet it may be absolutely wrong for him to sacrifice his parent to his own interests, in virtue of the special relationship between them. (Suppose there were no prohibition on his sacrificing himself to the greatest good of the state *when it is not his goal to do so.* Then if he must subordinate his own welfare to his parent's but not subordinate the interests of the state to her interests, he could be required to make the *sacrifice-for-her* of sacrificing himself rather then her to the state. But the state could not dominate his importance in any other case.)

In the same way *D* could dominate *P* through *P* dominates S^* and S^* dominates *D*. Likewise, if there were two duties, D_1 and D_2, it would be possible for $P > D_1$, $D_1 > D_2$, and yet $D_2 > P$, in virtue of the special relation between D_2 and *P*. There could be three duties, D_1, D_2, and D_3, and $D_1 > D_2$, $D_2 > D_3$, and yet $D_3 > D_1$. (A case illustrating this: Suppose we have duties to give to charities *A*, *B*, and *C*. In conflicts, it may be that we ought to give to *A* rather than to *B* because it does better work and to *B* rather than to *C* because it does better work, but give to *C* rather than *A*, since *A* is the offspring of *C*, and it would be wrong to prefer the offspring to the parent, even when the offspring does somewhat better work.) If *P* does not dominate *S*, but we have a choice between *P* and *S*, the same strong prohibition on sacrificing *D* to *P* can hold, regardless of what else we may sacrifice *D* to and what else may be sacrificed to *P*.

What the discussion of cases of strict dominance has shown is that, in addition to the explanation of intransitivities given above (132/3), we may sometimes fail to do our duty even though the objective value of what will be lost if we omit the supererogatory is *not* greater than the cost required for doing our duty. (For example, the state need not be seen as more important than the king, for the king to sacrifice his parent supererogatorily to the preservation of the state, even though he would have to sacrifice his own life for his parent.)

I conclude that the fact that we may sometimes violate a duty to do over-all good, and act for self-interest rather than over-all good, does not show we may violate a duty for self-interest, and the fact that we needn't make large efforts to pursue over-all good, but may pursue over-all good instead of duty, does not mean that we needn't make large efforts for the sake of duty.[13]

FRANCES MYRNA KAMM

New York University

---

[13] The intransitivities I have emphasized in this article are dependent on accepting the claim that supererogatory acts may sometimes take precedence over duties. But it is interesting to observe that those who do not accept this claim may be involved in an intransitivity of their own, at least when they also accept the view that one may sometimes make greater personal sacrifices in order to do a supererogatory act than one is required or willing to make to do one's duty. (We might refer to this as The Dutiful Person's Intransitivity.)

Suppose someone never fails to do his duty in order to do a supererogatory act, and makes all the efforts that are strictly required of him, *but not more,* in order to do his duty. He also does a supererogatory act at great personal sacrifice when it doesn't interfere with duty. The question arises why, if he is *willing* to make this great sacrifice to perform a beneficent act, consistency does not also require him to be willing to make the same sacrifice if it will help him to perform a duty which he thinks takes precedence over the supererogation. (A specific example: Someone will not leave work at his office to save a life, but he will not give up his life to finish his work at the office. In his spare time, he will cheerfully give up his life to save a life. If he considers his office work more important than saving a life, why won't he give up his life to do his office work, if he will give up his life to save a life?)

Accepting the view that there is an inconsistency in the dutiful person's position would imply that one could not consistently do all that is generally required of one to do one's duty, as well as forego a supererogatory act, and then devote even greater energies to doing nonobligatory beneficent acts—at least not if making these greater efforts could help one succeed in doing one's duty. Can one consistently think that doing the duty is worth omitting the supererogatory, but that the supererogatory is worth more effort than the duty?

I believe the dutiful person's position is not inconsistent—and so one cannot undermine his commitment to duty over supererogation by charging him with inconsistency. Once one has given duty what is owed it, one may do what one likes, without this implying that duty is owed more. This is because the factor that causes a duty to take precedence over supererogation may be different from the factor that makes it seem worth while to make a great sacrifice in order to do a nonobligatory good deed.

After this paper was written; Derek Parfit suggested to me that there is a connection between my claims that "make take precedence over," "dominates," and "is more important than," are not transitive, on the one hand, and, on the other hand, the claim of Ronald Dworkin and Amartya Sen that the relation "worse than" is not transitive and allows for only partial comparability. Parfit discusses this Dworkin-Sen claim in *Reasons and Persons* (New York: Oxford, 1984), pp. 430-432.

COMMENTS AND CRITICISM

MORAL DILEMMAS, DELIBERATION, AND CHOICE*

RUTH Barcan Marcus has recently[1] defended three theses regarding the nature of moral dilemmas. First,

> (M1) The very existence of moral dilemmas is evidence, not of intrinsic inconsistency in a moral system containing two principles that recommend incompatible actions, but rather of the pragmatic impossibility of acting by both those principles.

Her pragmatic corollary to M1 is a Kantian "regulative" principle regarding the resolution—or, more exactly, avoidance—of dilemmas:

> (M2) We ought to conduct our lives and arrange our institutions so as to minimize predicaments of moral conflict (121).

M2 is essentially telic, not deontic: we must, for prudential rather than specifically moral reasons, somehow order those (moral) principles in order to preclude as many situations of dilemma as possible. All moral principles are comparable; M2 holds irrespective of the categorical or conditional status of any two potentially conflicting principles. Yet not all principles are equal; life is unlike a card game in that we can "stack the deck" (130) to avoid conflicts.

M2, I will argue, is not simply an admission that we often resort to telic means to resolve dilemmas that are themselves deontic in nature. M2 really assumes not just that any moral principle can be a *member* of some ordering or hierarchy, but also that moral principles *are* hierarchical, i.e., "rationalized" preferences of one goal

---

*For their detailed commentary on earlier versions of this paper, I am indebted to Rulon Wells, Gordon Welty, and especially Terrance McConnell, who responded to parts I—IV at the University of Dayton Colloquium on Practical Reasoning (March 1983). Part VII was helped by Noam Chomsky's political works and correspondence. Sharon Anderson-Gold and Frederick Stoutland have straightened the tortuous path to my conclusions.

[1] "Moral Dilemmas and Consistency," this JOURNAL, LXXVII, 3 (March 1980): 121-136; parenthetical page references to Marcus will be to this paper. This continues a deontic tradition including: E. J. Lemmon, "Moral Dilemmas," *Philosophical Review*, LXXI, 1 (January 1962): 139-158; Bernard Williams, "Ethical Consistency," in his *Problems of the Self* (New York: Cambridge, 1973); Bas van Fraassen, "Values and the Heart's Command," this JOURNAL, LXX, 1 (January 1973): 5-19; Terrance McConnell, "Moral Dilemmas and Consistency in Ethics," *Canadian Journal of Philosophy*, VII, 2 (June 1978): 269-288; and Earl Conee, "Against Moral Dilemmas," *Philosophical Review*, XCI, (January 1982): 87-97.

0022-362X/85/8203/0139$02.40

over another. I will define and defend this claim by way of a critique of Marcus's third (Kantian) thesis, which assumes a deontic *definition* of specifically moral principles:

> (M3) Only normatively and formally universalizable choices, as resolutions of moral dilemmas, represent a *moral* preference of one to the other choice.

Marcus's case of the disagreement between E. M. Forster and his hypothetical interlocutor Worster, who respectively prefer "friend" and "country" when forced to choose, is alleged as evidence for the deontic claim that there is no *moral* ground for either preference: "there was disagreement . . . not about moral principles but rather about the kind of persons they wished to be" (135). This distinction between moral (universal) and personal (individual) preferences, I will argue, simply begs the question in favor of a deontic definition of 'moral': it forces the formalist hypothesis that there is some abstract way of telling whether the Forster/Worster disagreement is *either* moral or individual.

I do not take M1 and M2 to be inconsistent. Moral dilemmas may be deontic, though their resolution is telic. Moral theory, consequently, may be more of a patchwork business than we like. But if M1 and M2 are both *true*, then, given M3, the "regulative" principle in M2 cannot be a moral principle. M3 requires universality. But one could not, on Marcus's grounds, assume that in a pragmatic conflict between moral principles there is only *one* such regulative arrangement, and thus no *second*-order dilemma between possible regulative arrangements. Nor, assuming such a second-order dilemma, could one formally universalize the preference of one to the other arrangement. Marcus in fact sees us all in the same predicament as Forster and Worster; nobody can say *how* to regulate in a specifically moral way. But this result is unacceptable; hence the necessity of replacing M1–M3.

It is my basic aim here to reject the Kantian separation of moral and personal matters and to argue for an Aristotelian overlap between them. I will defend the Aristotelian notion that some reference to human *character* is necessary to the explanation of all human, and *a fortiori* all moral, action. Forster and Worster have a genuine moral disagreement. Moreover, their disagreement can be resolved only in terms of their respective moral choices about the "kind" of person each wants to be. In sections I–IV I will outline this Aristotelian account of morality, using Marcus's essay as a Kantian foil. In sections V–VII, I will pursue some implications of this account.

I

On Kant's categorical imperative, moral principles are formally universalizable, i.e., impersonal, maxims. On the alternative Aristotelian conception, moral principles are necessarily personal: only persons can exercise choice (*proairesis*) which, as its cognate 'to prefer' (*hairesthai*) implies, means the deliberative preference of one to another goal. Kant sees (apparent) moral dilemmas as mere occasions for doing our duty, but, as Marcus notes, insofar as only one duty stated by a dilemma can be done, guilt is the inevitable result of failing to do the other (130–133). But this is unacceptable, and requires that we construe moral maxims as personal: moral agents deserve to be guilty only for failing to keep maxims to which one would *rationally* subscribe. M2 actually assumes that there are situations where one cannot simply act on *a* (Kantian) maxim: one must in such cases *rationalize* the preference of one maxim over another. To the extent that M2 has this telic intent, I find it appropriate to search for *some* regulative principle that has the purpose of helping to resolve dilemmas. Moral agents need all the help they can get. To the extent that Marcus assumes that guilt is inevitable in situations of dilemma, i.e., that both maxims are morally binding, M2 interprets this "resolution" as the simple avoidance of dilemmas. In what follows, I wish to preserve the regulative function of M2, but to reject its Kantian presuppositions about guilt and avoidance. I will show how we can construe moral principles both as personal (as in Aristotle) and as binding (as in Kant), in order to construct an Aristotelian alternative to M2 which is not simply pragmatic but also moral. One of the ironies of M1–M3, noted above, is that their conjunction allows second-order dilemmas between different *pragmatic* principles for avoiding (first-order) dilemmas. We need an alternative to M2 that precludes second-order dilemmas.

Choice, in the Aristotelian sense, has two distinctive properties that ought to fulfill the above "binding" condition necessary to specifically moral principles. First, although it is necessarily a matter of personal preference, moral choice differs from aesthetic choice (or taste) in that it has deliberative regard for the consequences of action. Merely aesthetic choice stands in need of no rationalization. One simply prefers apple pie to chocolate cake; that is, there would be no moral consequence if this preference changed, or failed to bind. This is not to say that a given choice of preference cannot be both aesthetic and moral. One might prefer coffee to tea, while knowing also that tea is better for one's health, or being unable to buy coffee from a country that one would prefer to support politically. It is not on formal but on material grounds that we can distinguish between aesthetic and moral choice. If (a) a choice is made without regard to extrinsic or long-term consequences, and (b) there are in fact no such consequences, then that choice is aesthetic. The taste of apple pie is an intrinsic or short-term consequence of the aesthetic choice of one *object* over another. But the long-term effects of sugar consumption make the choice of one *category* over another—dessert over no dessert—a moral matter. Aristotle's assertion that "the lower animals cannot be said to be immoral [*akrate,* incontinent] if only because they lack the power of forming universal concepts [*kathalou,* i.e., categories]"[2] is a way of distinguishing between individual and universal concepts as a definitive feature of moral choice. Aristotle's occasional definition of moral choice as "deliberative desire" (NE 1113 a 11) means that there is a concrete and conscious process by which humans separate their immediate (aesthetic) from their long-term (moral) desires. The object of ethics is to "define the right choice"[3] regarding those actions which have more than intrinsic or immediate consequences.

A second "binding" property of moral choice is that, even when rational, it is not thought of as moral irrespective of persons and contexts of action. For example, Shakespeare describes Brutus and Cassius as exercising different choices regarding the same act of assassinating Caesar. Brutus chose the welfare of the state over personal friendship with a tyrannical ruler. Cassius chose personal advancement, though disguised as having the same lofty motivation. By the Kantian (formal) maxim that murder is categorically wrong, Brutus has no excuse. But Aristotle holds that assassination is wrong only if the relevant political structure is itself right: the context of a murder can make it conditionally right (or justifiable) rather than unconditionally wrong (or inexcusable). We justify Brutus, on Shakespeare's portrayal, because he acted on the long-term motive to eliminate tyranny; we blame Cassius because his motivations were only short-term.[4]

[2] *Nicomachean Ethics* (hereafter NE) 1147 b 4-5. I have generally adhered to Rackham's translation (Loeb ed.) except where consistency of vocabulary or a philosophical point demands clarification. For reasons that will become clear, I have generalized from "incontinent" to "immoral" here.

[3] *Protagoras* 356 e 5-6, regarding the "art of measurement."

[4] Cf. *Politics* 1267 a 13-16: "Since the greatest injuries spring from a desire for superfluities, not mere necessities, men do not for example become tyrants in order to avoid shivering with cold, and accordingly great honors are accorded not to one who kills a thief but to one who kills a tyrant."

Whereas the Kantian notion of universalizability takes the anticipated consequences of action to be irrelevant to morality,[5] what I will call *Aristotelian universalizability* takes such consequences to be essential. My argument against M3 concerns not whether agents can universalize maxims, but on what grounds they do. I take the Brutus/Cassius contrast to show that contexts and consequences are essential to deciding the morality of choice, not, as M3 implies, irrelevant to it. Aristotelian universalizability is much more restricted in scope than its formalist alternative in M3, however. To explain why Brutus' choice is morally universalizable requires the description of a specific context of choice, not a formal maxim. To speak of universalizability is to be able to judge another context as relevantly similar; in the plot against Hitler, Bonhoefer could have appealed to Brutus' but not to Cassius' example.[6]

On Aristotelian grounds, we could agree with Marcus's claim that "Not all questions of value are moral questions" (136) only for those questions which we have defined above as aesthetic. There is no explaining taste, or aesthetic values. But what Marcus calls matters of "sensibilities" may go well beyond the aesthetic and involve moral choice. Forster's and Worster's disagreement about the relative value of friend and country is a case in point. Even to take their preferences formally, that is, without reference to the particular friend and country involved, is to beg the question about the agent's abstract ability to universalize a given preference. Could Forster's sensibilities not include a case where some friend was simply not worth defending over his country? This very possibility shows that it would be morally *wrong* to universalize formally the category "friend" as a matter of sensibility; wrong, moreover, because it would fail to take into account the varying degrees to which one could value different friends or countries. By M3, however, we could not say "wrong" here; Marcus's formalist hypothesis holds that morality has "run out" at the horns of a dilemma, and this accounts for her amoral view of Forster and Worster. In fact, her very view of what a dilemma is calls for some Aristotelian commentary.

---

[5] Although contemporary Kantians have argued that Kant was not logically committed to this rigid view, I will take the deontic and telic views as ideal types (as M3 suggests) in order to show just how much of Kant's view is telic, either directly or in its presuppositions.

[6] Of Caesar, Brutus explains his act of assassination—it was "not that I loved Caesar less, but that I loved Rome more" (III,ii,32-37; cf. III,ii,224,181-183, 164-172; III,i,89; II,iii,127-129; II,i,166-180, 132-133, 116-124. And finally, "It must be by Caesar's death and for my part/ I know no personal cause to spurn at him/ But for the general"—II,i,10-12).

II

I will discuss elsewhere the exegetical issues surrounding the so-called "practical syllogism," Aristotle's official but sometimes cryptic account of deliberation.[7] Here, I will recast the clearer parts of that account, by assuming the isomorphism of (personal) deliberation and (interpersonal) conversation.[8] For present purposes, a theory of deliberation must state *how* an agent's choices or values are related to the resolution of moral dilemmas. Conforming to the logical principle of dilemma, I will not follow the Kant/Marcus assumption that there is a *conjunction* of moral principles competing for an agent's deliberative attention. Moral dilemmas have the disjunctive form:

(*D*) Should I pursue object(ive) *x*, or not?

The "not" option need not involve the pursuit of some alternative objective *y*; so-called "negative" actions regarding *x* could be construed without reference to any *y*. A relevant short-term interest may, for example, prompt the pursuit of some *x*, and a long-term interest, its avoidance.

Let me next contrast the deontic and telic concepts of moral dilemmas. Deontically conceived, to admit the reality of moral dilemmas is to acknowledge a *conjunction* of moral principles, not all of which can be satisfied. The disjunctive form *D* is thus alleged to reflect only a looser sense of 'dilemma' in which agents are wondering *how* to resolve difficult practical problems. The deontic viewpoint equivocates between these two senses: in the first (philosophical) sense dilemmas are ontic; in the looser (practical) sense they are merely epistemic. Philosophers who have not admitted ontic dilemmas have still admitted epistemic dilemmas.[9] The disjunctive viewpoint does not so equivocate. In fact, the main point of Aristotle's theory of the practical syllogism was to reconcile the ontic and epistemic notions of dilemma, in order to define what we mean by calling them *moral* dilemmas. I will thus turn to those features of his theory which are relevant to a critique of the deontic theory.

---

[7] Aristotle scholars will immediately recognize that my rendition is in basic sympathy with the "two-syllogism" reading of Aristotle's account, which holds that there are deliberative processes peculiar to *both* "reason" and "desire" and that the agent then acts for or against reason, that is, in a "restrained" or "unrestrained" manner. The practical syllogism was invented to explain such unrestraint or incontinence regarding reason, in answer to the question how it is possible both to have and to act against some moral principle. The "one-syllogism" reading holds that the only deliberative process involved is peculiar to reason, and that desires are direct causal (nondeliberative) factors that literally "pull" agents away from reason; consequently, deliberation does not provide a universal explanation of action. Given the notorious inadequacy of Aristotle's examples to his over-all theory, however, I have very deliberately distinguished between his notions of choice and what I have called the "ultimatum" of deliberation, sc. "the last premise, that originates action [*kuria ton prakseon*]" (NE 1147 b 10), in order to make his examples fit his theory of deliberation. An incontinent agent acts *with* an ultimatum, but *against* a (moral) choice, on the two-syllogism account of deliberation.

[8] See H. P. Grice, "Logic and Conversation," in Milton Munitz and Peter Unger, eds., *Semantics and Philosophy* (New York: NYU Press, 1975); Robert Stalnaker, "Pragmatic Presuppositions," in *ibid.*, and "Presuppositions," *Journal of Philosophical Logic*, II, 4 (October 1973): 447–457; and Richard Jeffrey, see below fn 11.

III

Any theory of deliberation about action must state the moral import of any possible discrepancies between the anticipated short- and long-term results of pursuing some desired object x. I will recast what Aristotle calls the moral choice of the *rational* deliberating agent as the *normative* value:

(V) x would produce pleasure (now) but unhappiness (finally).

It would be rational, given a belief in V, to avoid x.

For present purposes, the particular content of such a choice or value is irrelevant. Short- and long-term interests have been variously conceived by different individuals and cultures. Yet such a *rational* choice is binding or normative, not simply subjective, in that it is the frame of reference for all deliberation or conversation about what the individual or culture values. To admit that one "knew better" than to pursue x is to provide some interpretation of the relevant short- and long-term interests. Despite the many empirical differences between languages and cultures, however, we can isolate the *form* of any expressed choice of value, thus:

p. (x would produce pleasure now)
R. (but)
q. (x would produce unhappiness later)

This English expression, as an object of deliberation, has the general logical form qRp, where p represents a particular judgment of pleasure or short-term interest, q represents a general judgment of long-term interest, and R represents a relation of (rational) preference. Individually or absolutely considered, p and q represent the respective *reasons* for the pursuit and the avoidance of x. But R introduces the comparative realization that p and q cannot be considered absolutely. Like the word 'but' in this context, any substitution instance of R must meet the following conditions:[10]

---

[9] See McConnell, *op. cit.*, and Conee, *op. cit.* (in fn 1 above).

(1) It must denote that both p and q are believed true.
(2) It must denote that q is valued more than p.

Function (1) is the expression of conjunction; (2), of preference. Since p and q, however, are sentences that themselves express reasons for *contrary* actions, R cannot represent a two-place predicate that expresses the mere preference of some object (x) over another (y). Condition (1) implies that R must relate sentences rather than objects, and hence construes a theory of deliberation as an *explanation* of action. Only sentences can give reasons for action, and only deliberation can provide those sentences. Assuming (1), then, condition (2) requires that any substitution instance of R have the logical or syntactic status of a sentence connective, not of a two-place predicate. Expressions of preference, in a theory of deliberation, are expressions of relations between sentences that themselves express reasons for action.[11]

These remarks about the logical structure (qRp) of a rational choice or value (V) will help us to reconstruct an Aristotelian notion of moral principles that avoids the Kantian problems of M1–M3. First, we take sentences p and q above as minor premises, and as their respective major premises the two tenets of ethical hedonism:

P. Pleasant object(ive)s are to be pursued.
Q. Unpleasant object(ive)s are to be avoided.

Then we can represent the two different practical syllogisms, $(P, p \vdash Px)$ and $(Q, q \vdash {\sim}Px)$, where 'Px' means to pursue x and '${\sim}Px$' means to avoid x. But, so far, these two different deductive relations fail to incorporate our above remarks about the role of values or choices and, hence, of the relation R.[12] In this regard, I will adapt a remark from Aristotle's *Poetics* that it is not just an agent's "thought of what is best" but also his *character* that determines action (1450 a 1 ff.). By *presupposing* these two hedonist tenets, then, we can actually represent *four* different deliberative processes. As they define the deliberations peculiar to four different types of moral characters or agents, I will refer to the following deliberative schemata as *Aristotle's moral typology:*

---

[10] See NE 1144 a 31ff: "Deductive inferences [*syllogismoi*] about matters of conduct always have a major premise of the form 'Since the end or supreme good is so and so' (whatever it may be, since we may take it as anything we like for the sake of the argument)."

[11] As has been argued by Richard Jeffrey in "Preference among Preferences," this JOURNAL, LXXI, 13 (July 14, 1974): 377–391.

[12] The fact that there are two syllogisms, and two conclusions, shows that the standard two-syllogism account of Aristotle's theory of deliberation is inadequate. These two syllogisms account for only the incontinent and the temperate agents, whose respective beliefs p and q could be considered minor premises of syllogisms.

Moral types	First-order choice ("thought)	Second-order choice ("character")	Ultimatum of deliberation[13]
profligate	pRq ("his choice is unsound")[14]	pRq	Px ("Pursue x.")
incontinent (impetuous or weak)[15]	qRp	p ("he acts against choice")[16]	Px
continent	qRp	qRp ("he abides by choice")[17]	$\sim Px$ ("Avoid x.")
temperate	qRp	q ("he does not have evil desires")[16]	$\sim Px$

If we assume that qRp represents the normative value or *first*-order choice of the rational agent, we can define the notion of *second*-order choice to generate Aristotle's typology of moral agents. Equivalently, this is a way of drawing Aristotle's above distinction between thought and character as the two determinants of action. qRp, as an object of further deliberation, would be presupposed by any of three "typical" second-order choices. As a conjunction, qRp may be simplified in two possible ways on the occasion of some dilemma. In fact, the virtuous agent simplifies it to q, considering only the relevant long-term interest. I will thus interpret Aristotle's so-called "intellectualism," the claim that short-term interests ought to be considered irrelevant to action, as Aristotelian universalizability: we *ought* in this way to be virtuous. By contrast, the incontinent agent simplifies qRp to p, considering only the relevant short-term interest.[16] Finally, the continent or semi-virtuous agent *identifies* with qRp: in the literal sense that he draws an identity inference from that original value, he "abides by the conclusion he has reached" (NE 1145 b II). His thought and his character (or perhaps his "mind" and "spirit" in a religious sense) are "at one" (1166 a 13-14). More generally, our first-order choices or values identify our preferred *goals,* whereas our second-order choices identify *ourselves* as acting upon different possible deliberative attitudes toward those goals. A helpful analogy to these two orders of choice is Cardinal Newman's distinction[18] between types of assent. It is a notional form of assent to believe (e.g.) that cigarette smoking is wrong, but a real form of assent not to smoke: the real presupposes the notional form of assent.

---

[13] See fn 7 above. [14] NE 1152 a 15 ff.

[15] In NE 1150 b 19–28 (see also 1147 a 11–b 19), Aristotle distinguishes between the "weak" who deliberate and the "impetuous" who do not, before acting against their first-order choice. Because he regards the knowledge of ends (and thus, of choice between ends) as a condition of voluntary action, and voluntariness as a condition of moral responsibility, Aristotle does not seem to regard the impetuous agent as blameworthy. Our distinction between two orders of choice allows us to say, however, that it is possible for the impetuous agent to discount the long-term interest *before* the relevant perception of some desired object x. Such an agent is knowledgeable (and thus blameworthy) *when* making this choice, whether or not he is conscious of that choice when perceiving x: one can be ignorant because of a prior choice that was knowledgeable.

[16] "The temperate man cannot be continent, nor [conversely], for the temperate man does not have excessive or evil desires" (NE 1146 a 10 ff.; see 1104 b 4 ff., 1152 a 2–4). The incontinent agent "errs willingly . . . yet he is not wicked, for his moral choice is sound" (NE 1152 a 15 ff.; see 1166 b 7 ff.).

[17] NE 1145 b 11.

I will thus pin-point the philosophical alternative that this notion of a moral typology offers to the formalist definition of moral principles assumed in M3. Ever since Kant's distinction between prudential and categorical imperatives, it has been assumed that there is nothing peculiarly moral about a theory of deliberation. "Since I want goal x, I ought to employ means m" is a mere counsel of prudence. Incontinent action is taken to be a mere failure of prudence, and, *a fortiori*, a theory that describes or explains incontinence is taken to have no moral implications. These assumptions, however, beg the same questions as M3. It must be demonstrated, not assumed, that only formally universalizable choices are moral choices, that what one "wants" cannot be both prudent and moral. A study of the above typology can show how to accommodate a distinction between morality and prudence, but without assuming that they are necessarily exclusive domains.

It may first be noted that only the profligate disagrees about what ought to be desired. He, like Plato's democrat, simply "pursues the pleasure that offers," without that "view of the future" which is a criterion of rationality. "His choice is unsound" (NE 1152 a 15 ff), meaning that he cannot rationalize that choice in terms of its future consequences. But he *could* recommend his short-sighted maxims, the hedonism presupposed in major (universal) premises P and Q above, as formal maxims. Ethical hedonism is a normative position. It is on consequentalist grounds that the profligate's formal maxims are not universalizable.

More generally, this moral typology shows an alternative way to construe Marcus's distinction between morality and sensibility. If first-order choices define different moral preferences, we could thus interpret second-order choices as the (four) different types of sensibilities which moral characters could have *about* those first-order choices. Moreover, regarding the above Kantian problem about moral imperatives, only the virtuous or temperate agent's choice should be universalized. This is not to reject but to redraw Marcus's distinction between morality and sensibility. Sensibility presupposes morality; it does not simply concern a different range of choice.

---

[18] In *An Essay in Aid of a Grammar of Assent* (London: Longmans, 1901).

IV

It will aid my exposition of this moral typology to show how I construe second-order choice as definitive of different types of moral character. To do so, I will briefly discuss the philosophical antecedents of this distinction between orders of choice.

Harry Frankfurt introduced and defended the claim that there are orders of human *desire*.[19] Any self-moving beast, on his account, could have brute or first-order desires to do or to refrain from doing a particular act, whereas only a human being or person can have an attitude toward such desires. Only persons can have second-order desires, i.e., desires *to have* certain (first-order) desires or to "identify" with them in action. There are at least two problems with this view. First, as Frankfurt acknowledges, it has the consequence that his so-called "wanton" agent, who on that view simply fails to ask what his first-order desires ought to be, is not human. The question of the compositon of his second-order desires simply does not arise.

By comparing his wanton with Aristotle's profligate, as described above, we can see that Frankfurt simply begs the question about his wanton's humanity. This is not a terminological dispute, but a verificationist quandary, as can be shown by studying more entailments of the foregoing moral typology. The issue between my account of orders and Frankfurt's is whether the notion of choice has explanatory value. It *is* a choice, namely the choice to identify with his "unsound" preference of short- to long-term interest, that identifies the profligate as a particular character type. It is because of this particular first-order choice that Aristotle tentatively suggests that it is actually easier to convert the profligate than the incontinent agent to virtue: the former needs only to change his mind regarding "what is best" whereas the latter has the deeper personal problem of not identifying with his own choice of the best (NE 1146 a 31-b2). Aristotle later admits that this tentative suggestion was overly optimistic, in that incontinence is a *necessary step* on the profligate's road to virtue—the profligate must first change his mind to the "sound" choice qRp. But this is a different principle with which to identify, and thus a different personal problem (NE 1150 a 16-31; b29-36). Or, as Richard Jeffrey puts it (*op. cit.* 387), virtuous action is *not optional* for the profligate; the structure of the profligate's preference is such that he cannot, at least immediately, be virtuous. That the profligate ought to be virtuous does not presuppose that he can be virtuous. This is of course the same point that Frankfurt attempts to make, but because he simply *defines* the profligate (wanton) as one who has no desire for virtue he cannot offer a convincing proof.

---

[19] "Freedom of the Will and the Concept of a Person," this JOURNAL, LXVIII, 1 (Jan. 14, 1971): 5-20.

More generally, Aristotle's moral typology shows that it is not a truism to say that what an agent does, in resolving a dilemma, depends not only upon what he thinks but also on who he *is*, as evidenced by how that thought is regarded. Otherwise, moral theory would be back at square 1, the Socratic paradox that knowledge is sufficient for virtue. Or, to address an earlier problem in Marcus again: moral preference *and* a sensibility toward that preference are jointly sufficient for moral action, and *a fortiori* a particular preference and a particular character are jointly sufficient for virtue. As in tragedy (*Poetics*), so also in life (*Ethics*), thought and character are the two determinants of action.[20] It is essentially because of her Kantian formalism in M3 that Marcus recognizes only the "thought" determinant as constitutive of morality, and she thus consigns the "character" determinant to the amoral domain of "sensibility."

Gary Watson[21] cites a second problem for Frankfurt: for any desire, of any order, it is possible to be wanton or profligate with regard to that desire. In answer to this charge, and as a way of distinguishing between Frankfurt's theory and mine, I will conclude this reconstruction of the Aristotelian account of moral principles with a few general remarks. First, it is not the notion of orders that properly invites Watson's regress argument, but rather that of desires. Because an agent can have a theoretically infinite number of desires, Watson is right that there is nothing *about* a desire—for instance, its being second- rather than first-order—that gives desire an explanatory status regarding action. Yet this objection does not apply to the notion of choice. An agent cannot exercise an infinite number of choices regarding one action.[22] It was precisely because of the anti-Socratic claim that agents do not necessarily take their thought (or first-order choice) seriously that the notion of moral character (second-order choice) was invoked to explain how this is so. The pragmatic problem is that any such thought remains abstract, or untested, until concrete situations of dilemma between any two goals arise. It is in these situations that various "typical" second-order choices represent types of moral characater—but what role would there be for any third- and higher-order choice? Agents can literally identify with their first-order choices, in the simple sense that they can draw identity inferences from them. But they can only metaphorically identify with their desires; only figuratively can the whole identify with its proper part. Frankfurt's theory of desires invites Watson's regress charge because it fails to state what is at stake when agents do or do not identify. At stake is the whole issue of the moral identity of agents.

---

[20] This is the obvious analogy between *Poetics* VI and NE VII.

[21] "Free Agency," this JOURNAL, LXXII, 8 (April 24, 1975): 205-220, p. 218.

[22] Kierkegaard's theory of double reflection was formulated as a response to a similar "regress" problem in Hegel, the "bad infinite" in which reflection (deliberation)

To summarize these remarks about a moral typology and moral dilemmas, then: thesis M3 needs some consequentialist qualifications. There are choices that are not universalizable on consequentialist grounds, and thus not moral (e.g., ethical hedonism),[23] and choices that are thus universalizable and thereby moral (e.g., Brutus's). Having rejected the interdefinability of morality and formal universalizability, on consequentialist grounds we can replace M3 with

(A3) Moral principles are hierarchical: to act morally or immorally is to order one goal over another, in the primary (first-order) sense; it is to exhibit some sensibility toward that original ordering, in the secondary (second-order) sense. Only the virtuous agent's first- and second-order choices ought to be universalized.

Finally, contrary to the assumption in M1 that moral dilemmas are mere occasions for doing one duty, and *ipso facto* for failing to do another, I have argued for the corresponding telic definition of moral dilemmas, namely:

(A1) Moral dilemmas are occasions for ordering and reordering our preferences or ourselves.

---

might go on forever without ensuing in action. Kierkegaard held that there was both a reflection of understanding regarding the belief in some ethical principle and a reflection of the personality which exhibited an agent's mode of appropriating understanding—analogous to our two orders of choice.

[23] Although Kant's various injunctions against ethical hedonism have the appearance of deontic maxims, they all seem to be ways of spelling out his telic notion of "the ends of humanity." For example, that this end "in respect of sexuality is to preserve the species without debasing the person" [*Lectures on Ethics* (New York: Harper & Row, 1963), p. 170] derives maxims from a *telos*.

One may find some original ordering of one's goals irrational, in which case one ought to change it, or rational but painful, in which case the incontinent and continent agents adopt opposing attitudes toward it. Only the virtuous agent feels pleasure in constant *and* rational preferences; the rest of us are either pained by constancy (continent), inconstant (incontinent), or irrational (profligate) to begin with.

Enough regarding the nature of moral dilemmas and principles. In the remaining sections I will explore some entailments of the Aristotelian view, with particular emphasis on constructing an alternative to the regulative thesis M2, that is both pragmatic and moral and thereby precludes the apparent arbitrariness of regulative arrangements allowed by M2.

V

Marcus's various examples of moral dilemmas convey their intended philosophical morals only if thesis M3 is true. But the foregoing arguments, against M3's formalist notion of universalizability and its presupposition that dilemmas represent conjunctions of duties, need to be pursued more systematically. I will do this by applying our Aristotelian alternative A3 to Marcus's examples.

The first case on the continuum of examples is Marcus's rendition of the one-dimensional dilemma. Where two promises have been made, or two lives are at stake, and circumstances develop that force the agent to keep only one promise or save one life, are there any *moral* grounds for an otherwise arbitrary choice? Marcus argues that "however strong our wills and complete our knowledge we might be faced with a moral choice in which there are no moral grounds for favoring doing *x* over *y*" (125). This example is used as evidence against single-principle moral systems, such as Kant's categorical imperative (without the appeal to any regulative principle like M2): there is no single principle that can resolve such one-dimensional or symmetrical choices. But by virtue of what characteristic does such a symmetry obtain? Having rejected M3, we cannot cite any deontically (formally) defined maxim that would make options *x* and *y* symmetric. Without M3, it is up to an agent's consequentialist calculations to determine which option is more productive of long-term good, i.e., more moral. That is, there is no such thing as an a priori symmetric choice. It is a contingent truth that a particular dilemma would represent a symmetric choice; even if that choice were between two type-identical acts such as promise-keeping, the agent would need to know that both acts were equally disastrous.

A3 defines such one-dimensionality in terms not of formal maxims but of anticipated consequences. Whatever their respective rationales and sensibilities, Brutus and Cassius share the dilemma whether or not to kill Caesar. Each agent no doubt *anticipates* both personal and political consequences which are symmetric, that is, not a function of *who* commits the act. (Shakespeare at least invents no evidence that they were regarded differently by the Roman public.) Yet this consequentialist definition of one-dimensionality still allows us to say that Brutus acted morally, but Cassius did not. Brutus considered only the anticipated long-term consequences; Cassius, only the short-term.[24] Their respective sensibilities reveal two different moral characters. If, in deliberation at least, it is true that language mirrors reality, then Shakespeare shows how it is possible to act either morally or immorally in situations of symmetric choice. Brutus and Cassius, moreover, could be Forster and Worster.

VI

Marcus begins her discussion (121–123) of the nature of moral rules as the n-person case of personal rules that are definable simply as (first-order) choices in the Aristotelian sense. Moral dilemmas have general implications only when the n-person or "shared" principles of a positive moral code conflict in concrete situations; that is, only when it is impersonal codes rather than concrete persons who embody or experience the moral conflict. This is supposedly illustrated by the following case from Marcus:

> Plato . . . describes a case in which the return of a cache of arms has been promised to a man who, intent on mayhem, comes to claim them. Principles of promise keeping and benevolence generate conflict (121).[25]

If we are dealing with individual cases of moral conflict, however, we ought to name the individual parties rather than to treat the problem abstractly. I will thus transpose her example and then explore the ad hominem problems created by our foregoing reference to character as one of the two determinants of action. Suppose, then, that instead of the two abstract parties *A* and *B*, we have the concrete persons of Begin and Somoza as exemplars of the above moral dilemma. (Or make it Reagan and Rios-Montt; or, for that matter,

---

[24] Aristotle may seem to define "moral" here as exocentric action, and "immoral" as egocentric. Despite his aristocratic tendencies, however, he allows that insofar as exocentric action tends to legitimize institutions that are in fact unjust, it is morally wrong (see *Rhetoric* 1376 17–24; NE 1129 b 14–19; but, 1130 a).

[25] It is noteworthy that in Plato's own discussion of this example (*Republic* 331 c), Socrates says it would be *unjust* to return the arms.

Mondale and Marcos.)[26] Begin promised Somoza 10,000 Uzi submachine guns, has subsequently learned that Somoza is murdering his own people, etc. Although thesis M2 takes promise-keeping and benevolence to be comparable maxims—indeed, if they were not, there would be no problematic conjunction of (impersonal) duties—we have no reason to assume that these maxims have equal value. Moreover, we could accommodate the conceptual distinction that Kant draws between the perfect duty of promise-keeping and the imperfect duty of benevolence, as moral theorists: but to accommodate this distinction as moral agents is to order them.[27] What makes even a categorical duty "binding," in our previous terms, is the agent's preference to value it higher than some other duty as a matter of *both* first- and second-order choice. Begin, above, chose to make a promise binding;[28] whether or not such promise-keeping was rational as an original choice or value, he acted continently regarding it. But, just as in the case of quitting coffee drinking, where "inconstancy is in fact praiseworthy" regarding a choice one comes to regard as mistaken, constancy or continence regarding a mistaken choice can be regarded as blameworthy. It would be moral to change, rather than to adhere to or be bound by, a mistaken (first-order) choice.[29]

Marcus notes that the slogan, "You are damned if you do and you are damned if you don't," is coherent. It is perhaps also true, but it is not therefore a vindication of her claim that dilemmas represent conjunctions of (genuine) duties. Very typically, some parties will damn what others praise, regarding the particular choices an agent makes. I will explore the essentially *rhetorical* implications of this remark as an introduction to the final part of this paper, but will turn next to the more general issue it entails.

Ad hominem arguments, standardly conceived, fit into a philosophical view that rejects such "damning" remarks as irrelevant to any question of what an agent thinks. In our terms, it is wrong to assert that an agent's character determines his thought, and mistaken therefore to judge his character rather than the argument. But

[26] The first two political relationships are well documented. The last one was pointed out to me by Marvin Davidoff, of the Honeywell Project (Minneapolis): in 1977, Mondale placed a phone call to the White House, to reassure Marcos that his "crowd control" (napalm) was being delivered.

[27] Terrance McConnell has suggested that there might be an is-ought problem in my notion of "binding" duties. Promise-keeping ought to be binding, in moral terms, whether or not is *is* binding in some case. But my point is precisely that there are cases where promise-keeping ought *not* to be binding.

[28] See Noam Chomsky, *The Fateful Triangle* (Boston: South End Press, 1983), pp. 15, 25/6, 242; and *Towards a New Cold War* (New York: Pantheon, 1982), p. 16.

[29] NE 1146 a 17-19; 1151 b 4 ff.; *Rhetoric* 1376 b 17 ff.

our Aristotelian framework does not allege any such causal connection between character and thought. It is one type of thing, a first-order choice, to order one's goals; it is another type of thing, having the first as its deliberative object, to adopt an attitude toward that ordering. Thought and character, as types of choice, are the co-determinants of action. But character does not thereby determine thought. Thus an ad hominem argument can be construed as a critique of character, of the attitude or sensibility an agent has toward his or her values. Further, just as those values are deliberative objects for that agent, they are, more generally, possible objects of the public's proffered ad hominem critique.

With this analogy between the private (introspective) and public (critical) attitudes that are both possible with respect to agentive values, we can make a more general and constructive point about the ubiquity of ad hominem arguments. All moral dilemmas are occasions for such arguments. I will explicate this claim by applying it to the analysis of moral guilt. Marcus's analysis depends upon her presupposition that dilemmas represent conjunctions of duties: guilt is viewed as the inevitable result of the pragmatic inability to do both of two duties.[30] Her major illustration of this claim con-

---

[30] Marcus also makes this assumption of the essential parity of any two conflicting principles in her brief discussion of the Antigone case. Not only is Antigone obliged, by sororal duty, to bury her brother; Creon is also obliged to order her not to. I will offer both logical and exegetical critiques of this alleged conjunction of duties. First, if we interpret Oab as "a has an obligation to do b," and $Ckab$ as "the king morally commands a to do b," then Marcus is assuming:

(1) $$(x)(y)(Oxy \rightarrow Ckxy)$$

That is, an obligation y would exist for some x (e.g., Antigone) only if there were a command from k (Creon). But it is also true that Creon does not incur such an obligation; i.e., he does not *morally* command himself to do any y:

(2) $$(y) \sim Ckky$$

It is an obvious consequence of (1) and (2) that

(3) $$(y) \sim Oky$$

That is, Creon is not obligated to do any y. Unless, *per contra*, his obligations are definable as "divine rights" (as suggested by *Antigone* 11. 33–34: "[Creon] will be here to promulgate . . . his mandate [*to pragma agein*]").

But, even if it were independently assumed that Creon had a genuine obligation, there is no reason to consider it comparable to Antigone's obligation, as the deontic view would have us assume. Compare Aristotle's commentary (*Rhetoric* 1373 b 3–13; see 1375 a 25–63):

> Particular law is that which each community lays down and applies to its own members, this partly written and partly unwritten. Universal law is the law of nature. For there really is, as everyone to some extent divines, a natural justice and injustice that is binding on all men, even on those who have no association or covenant with each other. It is this that Sophocles' Antigone clearly means when she says that the burial of Polyneices was a just act in spite of the prohibition: she means it was just by nature:
>
> Not of today or yesterday it is,
> But lives eternal: none can date its birth.

cerns the necessary guilt of the woman confronting the dilemma of abortion.

Clearly, women disagree on the morality of abortion. It may be correct to view this disagreement as that between deontic and telic views, the former stating that abortion is categorically wrong and the latter that it can be conditionally right. Even if this were an accurate description of the dichotomy, however, it would fail to explain the status of the *subjective* guilt of the pregnant woman who *changed* her view from a negative to a positive evaluation of abortion. (I argued above that it can in fact be moral to change one's values, or first-order choices.) We may suppose the abortion argument to be between two persons at one time, or between two times for one person. Here, a woman comes to believe that abortion is objectively right, on whatever grounds, but she feels guilty nonetheless. She has the feeling, but not because of the fact, of guilt. Marcus's discussion has no place for this distinction between fact and feeling, just as her discussion of "ought" has no distinction between (universal) duties and (individual) values. Although the woman in our example changes her values regarding the morality of abortion, she hypothetically continues to travel the road not taken, to identify with her old values in imagination but not in action. And that is precisely why she feels guilty: she is "incontinent" with respect to that old road, and *could* not feel guilty if she had stayed on that road. Whether or not we want to, we all identify with our former values to the extent that we are historical creatures. More generally, we can be individually the victims of ad hominem arguments against ourselves. Subjective guilt might be inevitable, as a function of personal value changes, but guilt is objectively appropriate only if the new value is wrong.[31] It is when people stop identifying with the old and start identifying with their new values that they have a measure of moral growth—they stop feeling guilty for what they think is right.

VII

One side of the coin of assessing agentive responsibility is to blame—whether oneself or others. Its obverse side is to praise, and both activities have been traditionally conceived as the business of *rhetoric*. In closing this essay, I will elaborate upon the foregoing remark that there is in an important sense only a rhetorical difference between damning and praising some action. This elaboration will constitute a rationale for replacing thesis M2. Although we can all use regulative advice for resolving moral dilemmas, a general adequacy condition that M2 fails is that it create no second-order dilemmas regarding its own application. Further, M2 as stated begs an important question about the compatibility of individual and collective interests; that "we should arrange our lives and institutions" in optimal ways is to assume that there are no situations where the lessening of individual dilemmas *is* the multiplying of social dilemmas. What needs to be shown, in the attempt to replace M2, is the relation between rhetoric and morality in the description of choices.

---

[31] When Patricia Greenspan says that "whether guilt is at *all* appropriate depends on the nature of the countervailing reasons" ["Moral Dilemmas and Guilt," *Philosophical Studies* 43, 1983: p. 121], I take those reasons to be the substituends for q in the relevant value qRp.

There will inevitably be rhetorical tendencies in the very description of dilemmatic choices. It is of course a philosophical commonplace, since Donald Davidson made the point that causal laws connect not events but descriptions of events,[32] that there are important disagreements about the descriptions of action. The Aristotelian view developed here can acknowledge people's motivations for offering the action descriptions they do, while also addressing the regulative issue how actions ought to be described.

The best way to get at these phenomenological and normative issues is to offer some examples. I will start with a case from the personal sphere. Suppose, then, that one party to a marriage has, but is tempted by some desired person *x* to trespass, a rather standard value regarding the potential dilemma of an extramarital affair. Assuming that the marriage contract is a serious one for both marital parties, the relevant value that they share is presumably

(*V*) Person *x* would produce pleasure (gratification, personal power, etc.) now, but unhappiness (regret, discord, etc.) in the long run.

To shift my emphasis from an earlier point: the 'but' connecting these short- and long-term desiderata has two functions. Besides expressing either marital partner's rational preference regarding the matter at hand, it also has the more basic function of a logical '&'. A potential marital transgressor may thus choose to *simplify* this conjunction of desiderata in *V*. Logically speaking, acts of simplication are literally value-neutral. But rhetorically speaking, they are not. The agent, in some attempt to excuse the act, may describe this simplication of *V* as the (mere) *selection* of the short-term desideratum—roughly, and generally, defense as:

---

[32] "Mental Events," in L. Foster and J. W. Swanson, eds., *Experience and Theory* (Amherst: U of Massachusetts Press, 1970), pp. 80/1: "But laws are linguistic, and so events can instantiate laws . . . only as those events are described in one way or another."

(*S*) I just meant to enjoy myself; I did not mean to hurt you. (And, you should not take it so personally, etc.)

It is etymologically informative that Aristotle's incontinent agent, who is "the pursuer" (*ho diokon*) of *x*, is also "the avoider" (*ho pheugon*) of legal retributions for that act. And conversely, the legal plaintiff or accuser is also the pursuer of that retribution,[33] in Attic dialect. It is not surprising, then, that the offended marital party may describe that same choice of *x* as a plaintiff. An accuser or *analyst,* more generally, may seek to blame that act, by describing that same simplication of *V* as a *discounting* of the long-term interest, roughly as:

(*D*) You meant to destroy our marriage. (You tried to assert control, etc.)

What we could call such agent/analyst disputes, between descriptions *S* and *D,* are both more common and more rhetorical than most of us would like to admit. But more interesting logically is that *S* and *D* are *equivalent* descriptions, whatever the rhetorical intent of agent or analyst: to select the first conjunct of *V* is to discount the second. Descriptions *S* and *D* are not therefore psychologically or rhetorically equivalent, however; their logical equivalence is only a necessary condition for the agent's and analyst's disagreement about how to describe the *same* act. Only if agent and analyst share the value, *V,* or (if the agent is a profligate) both know what it would be to share *V,* would they be *able* to disagree about the same act *x*.

Descriptions *S* and *D*, of course, are only egocentric idealizations. A guilt-ridden or penitent agent may use description *D,* while a forgiving or enlightened analyst may use *S*. Or, a "weak" (nonimpetuous) incontinent agent may have complex motivations, including both *S* and *D*. These caveats are meant to anticipate a normative point about the rhetoric of act descriptions. Restricted to a purely rhetorical level, the agent/analyst dispute would look like thesis

(R) The agent describes the choice of *x* in terms of *S*, as his or her true intention. The analyst uses the terms of *D*, to impute a contrary intention as the true one. There is no court of outside appeal that can decide that true intention.

---

[33] See *Rhetoric* 1, 1-3; 13-15; 18-19.

But there is also a metaphysical level to their dispute about descriptions. Stated as a paraphrase of the medieval doctrine of double effects, this is the metaphysical or normative thesis

> (N) An agent intends all the consequential entailments of his or her actions, as those entailments are stated by the relevant value *V*, regardless of which consequences that agent says he or she intends.

Thesis N is an analogy to a standard epistemological view which holds that one knows the "obvious content" of a perceptual claim *p*.[34] There have been objections that this view cannot distinguish between what is obvious about *p* and what is not. With N, however, there is no such ambiguity: the actual content of the short- and long-term desiderata in *V* will determine the scope of these consequential entailments (or "obvious content"). In our earlier terms, an agent knows *what* he believes when interpreting some value *V* as the relation *qRp*. If there are *un*foreseen consequences of some choice *x*, they are *not* intended by choosing *x*. For instance, if an agent is originally ignorant concerning either the political or the personal health effects of coffee drinking, he does not intend those consequences by drinking that coffee. Only after knowing such effects would he be responsible for them as consequential entailments of coffee drinking.

By thesis R above, we can read the descriptions offered by agents and analysts as the necessary clues to their respective or shared values. By thesis N, we can in turn read agents' or analysts' values as necessary clues to the ethical question of agentive responsibility and, finally, to the metaphysical question of agentive freedom. I will reserve this metaphysical issue for another occasion, but will conclude with a discussion of the ethical issue.

What can be made of our anticipatory remark above that an agent may have complex motivations or intentions for action? Kant was wise enough to conclude from this possiblity that an agent, in principle, cannot *know* his or her true intentions (*op. cit.* 74/5). What Kant called empirical (or pathological) reasons, namely those of the body, may be parsed generally as short-term interests. What he called rational (or maxim-based) reasons may be parsed, by our foregoing anti-formalist arguments, as long-term interests. But Kant's agnosticism about *de facto* reasons does not preclude agentive knowledge of the possible reasons for doing some action *x*. An agent knows what would count as short- or long-term interests in a given context of choice, to the extent that they comprise his or her relevant values. All moral (deliberative) agents are still within this domain of considering the possible. It is, moreover, only for those who deliberate that there could be such a thing as a regulative principle about how to deliberate.

---

[34] Carl Ginet, *Knowledge, Perception and Memory* (Boston: Reidel, 1976), p. 9.

As a virtual paraphrase of a standard Kierkegaardian thesis, then, I will offer as an alternative to Marcus's regulative thesis M2 the thesis

> (A2) One ought to be objective, or descriptively disinterested, regarding one's own intentions, and subjective or interested regarding other person's intentions.

In our foregoing terms, A2 means that agents should become analysts, *and* conversely, regarding their descriptions of actions. As a response to M2, A2 is both rhetorical and ethical: it is about the sensibility that agents could or should have. That one *ought* to see actions as others do is just to say that matters of sensibility *are* moral matters. That is my final response to the alleged amorality of the sphere of personal preference.

But is A2 simply vacuous? Does it in fact regulate our behavior in substantive ways, or is it merely an existentialist preoccupation? I will argue the former option, by studying the following extended quotation from Noam Chomsky:[35]

> In fact, the concept "American aggression" is unthinkable, perhaps an oxymoron. One may speak of "the Vietnam War (which we entered out of fear of Chinese expansionism)" or describe our intervention as an unfortunate or tragic error but the well-documented facts have been effectively expunged from official history.
>
> To assess the performance of the free press in this regard it is sometimes useful to ask how we would react to similar pronouncements in the propaganda organs of our official enemies. Suppose, for example, that we were to read the following complaint in some communist party organ:
>
>> For a segment of Western Europe's right, Afghanistan is a windfall. Exotic, blessedly far from home and potentially confirming a post-Czechoslovakia generation's notions of the misuse of Soviet power, the controversy over the war is packed with possibilities for simplistic portrayal. If offers Europeans the luxuries of dis-

---

[35] Compare this with Ronald Reagan's admonition, to the National Association of Evangelicals (April 1983), to beware of the temptation "to declare yourselves above it all and label both sides equally at fault, to ignore the facts of history and the aggressive impulses of an evil empire, to call the arms race simply a giant misunderstanding and thereby remove yourselves from the struggle between right and wrong and good and evil." Reagan is continuing a long presidential policy, as I point out in "The Representation and Resolution of the Nuclear Conflict," in a "war" issue of *Philosophy and Social Criticism*, x, 3/4 (March 1985).

> tance and vagueness. Also, it drowns out discomfort about Vietnam, distracts attention from El Salvador, and tends to confirm fond suspicions that superpowers behave alike.
>
> (*Towards a New Cold War,* 12/13)

Chomsky continues, "Sickening, no doubt, but what can one expect from a controlled party press? In fact the passage is virtually a transposition of a passage on El Salvador in the *New York Times,* with the obvious changes of names." That passage goes:

> For a segment of Western Europe's left, El Salvador is a windfall. Exotic, blessedly far from home and potentially confirming a post-Vietnam generation's notions of the misuse of United States power, the controversy over the little war is packed with possibilities for simplistic portrayal. Also, it drowns out . . . discomfort about Afghanistan, distracts attention from Poland . . . and tends to confirm fond suspicions that superpowers behave alike.

An application of A2, not politics, is the immediate point of this lengthy quotation. The ethics of act descriptions, by A2, ought to regulate the rhetoric of those descriptions. To use one particular from Chomsky for an illustration of the consequences of an "unregulated" description, we might say that it is *because* "the concept 'American aggression' is unthinkable" to American editorial minds that "the well-documented facts have been effectively expunged from official history." Because those facts do not appear today in high school history books or classrooms, military registration is not posed as a moral problem for the average candidate for U.S. interventionist forces around the globe. A2 is fundamentally about "editorial minds" in the widest sense, about people who select only the information that they want to hear about contexts of action. The fundamental moral intuition behind A2 is that "one should do unto others as unto oneself," or should "love one's neighbor as oneself."

Thesis M2, on the other hand, is dangerously deficient on the question of just how "our lives and institutions" ought to be regulated in order to avoid dilemmas. I have already remarked about the question begged by the word 'and' in M2: it would allow the *de facto* regulators of any society to minimize their own predicaments of dilemma, but in a highly interdependent web of personal and political relations this simply *is* to increase others' predicaments.[36] With A2, however, there is some hope for instituting a genuine social or political order, without this second-order dilemma of who is to do the ordering and how. This is not to claim that A2 is a sufficient condition for solving the world's regulatory problems. A2, though hardly an original principle, is still an origin of all *principled* change.

---

[36] Robert Drinan has aptly located this problem regarding the dilemma between American affluence ("freedom") and Third World subsistence, in his speech title "When Moral Imperatives Collide, Who Shall Decide?" (Midwest Scholars Dialogue on Religion, Antioch College, November 1981).

LYLE V. ANDERSON

Columbus, Ohio

A SOLUTION TO FORRESTER'S PARADOX OF GENTLE MURDER*

PARADOXES both irritate and stimulate deontic logicians. The paradox of the good samaritan is probably the best-known example, but its traditional versions depend on scope confusions. Such confusions are supposed to be absent from a new version, called the *paradox of gentle murder,* which was presented recently by James Forrester.† I will argue that Forrester's paradox also depends on a scope confusion and that the paradox can be avoided by analyzing the logical form of the action sentences inside the scope of the deontic operator. Thus, Forrester's paradox reinforces the importance of logical form.

Forrester's paradox is roughly that, even if murders occur, it seems obligatory for the murders to be gentle, but, on standard deontic logic, if it is obligatory to murder gently, it is obligatory to murder, and that seems wrong. More precisely, Forrester presents his paradox by deriving a contradiction from a situation and plausible principles of deontic logic. The situation is specified by two legal rules and an action sentence:

(1) It is obligatory$_l$ (Read: obligatory under the law) that Smith not murder Jones.
(2) It is obligatory$_l$ that, if Smith murders Jones, Smith murder Jones gently.
(5) Smith murders Jones.

The first principle that Forrester assumes is

(3) $$[O_r(p \supset q) \supset . p \supset O_r q]$$

---

*Thanks for comments on earlier drafts are due to Bob Barnes, Bob Fogelin, and Lynne McFall.

†"Gentle Murder, or the Adverbial Samaritan," this JOURNAL, LXXXI, 4 (April 1984): 193-197. (1)-(10) are quoted from this article, and all references in the text are to this article.

0022-362X/85/8203/0162$00.70

where the subscript 'r' indicates that the obligation can derive from any source, including the law. (3) plus (2) yields

(4) If Smith murders Jones, it is obligatory$_l$ that Smith murder Jones gently.

(4) and (5) imply

(6) It is obligatory$_l$ that Smith murder Jones gently.

But it is necessary that

(7) If Smith murders Jones gently, then Smith murders Jones.

and standard deontic logics include a rule

(8) $$\frac{\vdash (p \supset q)}{\vdash (O_r p \supset O_r q)}$$

and (6), (7), and (8) imply

(9) It is obligatory$_l$ that Smith murder Jones.

The final principle that Forrester assumes is

(10) $$O_r p \supset {\sim}O_r{\sim}p$$

and (1), (9), and (10) yield the contradiction

(11) $$O_l{\sim}p \mathbin{\&} {\sim}O_l{\sim}p$$

The contradiction can be avoided only by rejecting at least one step of the argument. Forrester says, "steps (1), (2), and (5) are the basic facts of the case: there seems no reason to consider them mutually inconsistent" (195). (4), (6), (9), and (11) are validly inferred from previous steps. (7) can be supported by any plausible theory of adverbs. Forrester argues for (3) in this case, and he says "giving up (10) will not be enough" (196). Forrester concludes that (8) should be rejected, i.e., that truth is not always preserved when a logical implication of a sentence is substituted for that sentence inside the scope of a deontic operator.

This argument can be generalized. Forrester's example concerns legal obligations, but the crucial steps (3), (8), and (10) use the subscript 'r', which applies to any deontic operator; so similar arguments can be formulated for moral obligations and for what agents legally or morally ought to do. Thus, if Forrester's argument works, no deontic operator allows substitution *salva veritate* of logical implications.[1]

This conclusion should be very disturbing. Forrester sees that, if

---

[1] Because the argument applies to every deontic operator, I will henceforth drop subscripts.

(8) is rejected, most of standard deontic logic must be reconstructed, and this is "a large task" (197). Furthermore, many arguments that seem obviously valid could not be justified without some rule like (8). For example, if I both mow and water your grass, I mow your grass, so, if it is obligatory for me to mow and water your grass, it is obligatory for me to mow your grass. Such arguments cannot be justified if (8) is rejected, unless some other rule or principle is substituted. Forrester gives no substitute for (8) in his article, and it seems that any substitute would have to be very complex and unintuitive in order to justify all such obviously valid arguments.[2] Thus, it is preferable to avoid the paradox without denying (8).

One way to do so is to deny (3). Forrester does not defend (3) in all cases, but he does argue for (3) in the present case. His arguments are inconclusive,[3] but (3) still seems plausible in this case. Furthermore, even if (3) is denied, a contradiction still seems to be derivable from (1), (4), and (5), so some paradox remains. Consequently, I will grant (3) in this case for the sake of argument.[4]

Another way to avoid the paradox is to deny (10). (10) is denied on independent grounds by advocates of the possibility of legal and moral dilemmas. Instead of defending (10), Forrester says, "giving up (10) will not be enough, for even if we end the argument with step (9), the conclusion, although not a paradox, is pretty unwelcome" (196/7). However, (9) is not unwelcome if the situation in

---

[2] In "Aspectual Actions and the Deepest Paradox of Deontic Logic" (delivered at the Davidson Conference, Rutgers University, April 28, 1984), Hector-Neri Castañeda proposes a limited substitute for (8) which does escape Forrester's paradox. Castañeda's solution depends on a complicated ontology of various kinds of acts. However, such complications are unnecessary, since my solution works with only quantifiers and scope distinctions. In spite of such difficulties, I owe much to Castañeda's discussion.

[3] Forrester's arguments for (3) come down to the claims that it is plausible to infer 'Smith ought to kill Jones gently' from (2) and (5), but "only something like (3) can sanction such a move" (196). However, according to Davidson's theory (see below), 'Smith ought to kill Jones gently' can be interpreted as (6.2): '$(\exists x)(Mxsj \,\&\, OGx)$', and the move from (2) and (5) to (6.2) can be sanctioned by

(3.2) $O[(\exists x)Mxsj \supset (\exists x)(Mxsj \,\&\, Gx)] \supset [(\exists x)Mxsj \supset (\exists x)(Mxsj \,\&\, OGx)]$

Thus, Forrester's arguments fail to support (3) over (3.2). If (3) is replaced by (3.2), no contradiction can be derived from (1), (2), and (5).

[4] In his comments as referee for this JOURNAL, James Higginbotham showed that (3) and (8) lead to absurdities. If we put '$p \supset q$' for 'p' in (3), we get (i): '$O((p \supset q) \supset q) \supset ((p \supset q) \supset Oq)$'. But (ii): '$p \supset ((p \supset q) \supset q)$' is a tautology; so (8) yields (iii): '$Op \supset O((p \supset q) \supset q)$'. (i) and (iii) imply (iv): '$Op \supset ((p \supset q) \supset Oq)$' which implies (v): '$(Op \,\&\, \sim p) \supset Oq$'. This is absurd, since it means that, if anything is obligatory but is not done, then everything is obligatory. Thus, anyone who defends (8) must reject (3) as a general principle. Nonetheless, a restricted version of (3) might still warrant the inference from (2) to (4) in this case, as Forrester claims, and we still need to solve the paradox that (1), (4), and (5) seem to imply a contradiction.

Forrester's paradox is seen as a legal dilemma [relative to laws (1) and (2)]. Nonetheless, the situation does not seem to be a dilemma, since Smith can follow both laws by not murdering Jones. Consequently, (9) is unwelcome in this case, so some paradox remains even without (10).

The best solution to Forrester's paradox is not to deny (3), (8), or (10) but instead to claim that (1), (2), and (5) are jointly inconsistent. Forrester says only that "there seems no reason to consider them mutually inconsistent" (195). However, the reason to consider them inconsistent is that they imply a contradiction if otherwise acceptable principles are assumed. What better reason could you want?

Forrester would probably respond that (1), (2), and (5) seem consistent, so any principles of deontic logic that allow us to derive a contradiction do not accord with our common ways of speaking and reasoning with deontic operators. However, it is easy to explain why (1), (2), and (5) seem consistent when they really are not. We should suspect that the problem lies in (2), because (1) and (5) are more common than (2). If some other principle is similar to (2) but consistent with (1) and (5), then anyone who confuses this other principle with (2) will think that (2) is consistent with (1) and (5).

Which principle is confused with (2)? Forrester writes, "the legal system might have stated (4) as its principle in place of (2)" (194). However, the contradiction follows from (4), assuming (8) and (10), so (4) provides no escape.

(4) and (2) are the only alternatives Forrester considers, and the reason is clear. Forrester uses propositional constants and avoids quantifiers "to simplify matters" (194); so the deontic operator can have only two scopes. Its scope can be wide if it includes the whole conditional, as in (2), or narrow if it includes only the consequent, as in (4).

An even narrower scope becomes available if we analyze the logical form of the action sentences within the scope of the deontic operator. Following Donald Davidson's theory,[5] we can analyze (5) as

(5.1) There is an act of murder by Smith of Jones.

or symbolically as

(5.2) $(\exists x)Mxsj$[6]

---

[5] "The Logical Form of Action Sentences," in *Essays on Actions and Events* (New York: Oxford, 1980), pp. 105–122.

[6] (5.2) could be broken down further into '$(\exists x)(M^{*}x \;\&\; \mathrm{BY}(x,\mathrm{Smith}) \;\&\; \mathrm{OF}(x,\mathrm{Jones}))$', but a complete analysis is not necessary for my purposes.

The antecedent of (7)

(7a) Smith murders Jones gently.

can then be analyzed as

(7.1a) There is an act of murder by Smith of Jones and it is gentle.

or symbolically as

(7.2a) $(\exists x)(Mxsj \mathbin{\&} Gx)$

This analysis shows why (7) is true. More relevantly, it allows several interpretations of (6), because the deontic operator can apply at several places or take several scopes, including

(6.1) $O(\exists x)(Mxsj \mathbin{\&} Gx)$

(6.2) $(\exists x)(Mxsj \mathbin{\&} OGx)$

Roughly, (6.1) says the occurrence of an act is obligatory, whereas (6.2) says the act occurs and its character as gentle is obligatory.

These distinctions allow various interpretations of (4). Taken strictly, (4) can be interpreted with (6.1) as its consequent:

(4.1) $(\exists x)Mxsj \supset O(\exists x)(Mxsj \mathbin{\&} Gx)$

or with (6.2) as its consequent:

(4.2) $(\exists x)Mxsj \supset (\exists x)(Mxsj \mathbin{\&} OGx)$[7]

Since (4.1) and (4.2) are so similar, they are easily confused. However, (4.2) avoids the paradox.

No contradiction can be derived from (1), (4.2), and (5), because (6.1) does not follow from (4.2) and (5). All that follows from (4.2) and (5) is (6.2). However, if (7) is analyzed as

(7.2) $(\exists x)(Mxsj \mathbin{\&} Gx) \supset (\exists x)Mxsj$

then the rule (8) warrants only the conclusion

(8.1c) $O(\exists x)(Mxsj \mathbin{\&} Gx) \supset O(\exists x)Mxsj$

and the antecedent of (8.1c) is (6.1) rather than (6.2). Since (6.1) cannot be derived from (4.2) and (5), neither can (9). Therefore, (1), (4.2), and (5) are consistent. Their consistency plus an understandable scope confusion of (4.2) with (2) or (4.1) explains why (1), (2), and (5) seem consistent when they are not.

(4.2) also serves all the legitimate purposes of (2) or (4). Forrester says that the purpose of a law such as (2) or (4) is to show that the legal system "considers murdering violently to be a worse crime than murdering gently" (194). (4.2) serves this purpose as well as (2) or (4), because murdering violently violates (4.2), but murdering gently does not violate (4.2).

---

[7] Someone who says (4) probably means to say (4.3): '$O(x)(Mxsj \supset Gx)$' or (4.4): '$(x)(Mxsj \supset OGx)$', because these avoid the absurd implication of (4.1) that, if Smith murders Jones roughly, it is obligatory for Smith to return and murder him gently. In any case, my solution is unaffected, because the contradiction follows from (4.3) but not from (4.4).

Furthermore, the justification for (4.2) is simple and intuitive. Forrester defines "'doing A to person m at time t gently' as 'doing A to m at t in such a way as to cause m the least amount of pain consistent with the agent's doing A to m'" (194). If a legal system wants to make it obligatory to cause the least pain consistent with doing an act of any kind, then the legal system can include a general rule that

110154

(12) $(x)OGx$

which implies a general rule that, for any kind of act, A,

(12.1) $(x)\,(Ax \supset OGx)$

which implies that, for any kind of act, A,

(12.2) $(\exists x)Ax \supset (\exists x)(Ax \,\&\, OGx)$

(4.2) is an instance of (12.2); so (12)–(12.2) can justify (4.2). No general rationale for (2) or (4.1) is this plausible.

Thus, there are several reasons to interpret any actual law like (2) or (4) as (4.2) rather than as (4.1) and to prefer (4.2) to (2) or (4.1) as a moral rule. If anyone insists on (2) or (4.1) as a law or moral rule, and they also accept (1) and (5), then their views are inconsistent. Such inconsistency is not a paradox.

Critics might respond that this solution assumes Davidson's theory of the logical form of action sentences, but that theory is questionable, and there are alternative theories of adverbs and action sentences on which no scope solution to Forrester's paradox is available. The most plausible alternatives[8] analyze adverbs as operators on sentences, so

(7a) Smith murders Jones gently.

is analyzed as

(7a*) Gently (Smith murders Jones)

(6) must then be analyzed as

(6*) O(Gently (Smith murders Jones))

---

[8] Cf. Romane Clark, "Concerning the Logic of Predicate Modifiers," *Noûs*, IV, 4 (November 1970): 311–335, and Terence Parsons, "Some Problems concerning the Logic of Grammatical Modifiers," *Synthese*, XXI, 3 (October 1970): 320–334.

and the contradiction follows. Such operator analyses do not allow a scope where a deontic operator applies to the adverb 'gently' but not to the verb 'murders', so they do not allow any solution to Forrester's paradox except denying (3), (8), or (10). I cannot here discuss all the pros and cons of various approaches to adverbs. However, if a theory of adverbs or action sentences does not allow a plausible solution to the paradox of gentle murder, so much the worse for that theory. One advantage of Davidson's theory is that it solves the paradox of gentle murder.

A second possible objection is that the quantifier over acts in (4.2) lies outside the scope of the deontic operator, so the quantifier ranges over actual or performed actions, rather than over acts which are obligatory but which might not be done. This is admittedly unusual, but it does not cause any problems in the context of the paradox. The quantifier over actual acts occurs only in the consequent of a conditional or as the conclusion of an argument where the antecedent or premises guarantee that some act of murder is actual. Thus, the consequent of (4.2) will never stand alone unless there is an actual act for the quantifier to range over. If someone does judge that 'It is obligatory that Smith murder Jones gently' before Smith actually murders Jones, then the judger must be assuming that Smith will murder Jones, so the judgment is elliptical for a conditional judgment. In such cases, nothing is wrong with quantifying into deontic contexts.

In the absence of further objections, I conclude that Forrester's paradox depends on a scope confusion of (4.2) with (2) or (4.1). The moral is clear. Paradoxes can arise from inadequate attention to the logical form of action sentences inside the scope of deontic operators. In order to avoid such paradoxes, one must use not only propositional constants but also more complicated analyses where the logical form is explicit.

Dartmouth College WALTER SINNOTT-ARMSTRONG

13/11/44

110154

Compiled
1999-2000